Predicting Wine Club Attrition

Steve Bowden
MS Analytics 2015

Primary Research Objective

Situation = Winery A and Winery B have developed very profitable wine clubs. However, growth in this key area is constrained by member attrition averaging 25% per year. Unless this attrition rate can be lowered, an untapped source of new members can be identified or prices can be raised, the profits generated through these clubs will level off.

Complication = Currently there are no anti-attrition remedies being tested because we cannot accurately identify those members with a high likelihood of canceling their membership. Thus, we are forced to be reactive rather than proactive.

Question = Can we develop predictive models to assess the likelihood a wine club member will cancel their membership?

Initial Hypothesis (Answer) = Yes, using account-level, transaction-level and geographic indicators, it is possible to determine the likelihood that a member will cancel their membership.

MECE Diagram

What are the potential reasons for cancelling a Wine Club membership and can we identify proxies for customer behavior to help predict future cancellations?

Service Factors

- Customer Service
- Transit time too long
- Not home to sign
- Damaged shipment

Engagement Factors

- Too few/many emails
- Unsubscribes
- Email Bounces
- Outbound calls
- Event attendance
- Tasting room visits
- Non-club orders

Product Factors

- Don’t like the wine
- Have too much
- Too expensive
- Want more variety
- Want different wine
- Corked wine

Demographic Factors

- Age
- Income
- Length of Residence
- Moved?
- Presence of children
- Type of credit cards
- Gender
- Home Ownership
- Adults in Household

Geographic Factors

- Zip-level census data
- Miles from winery

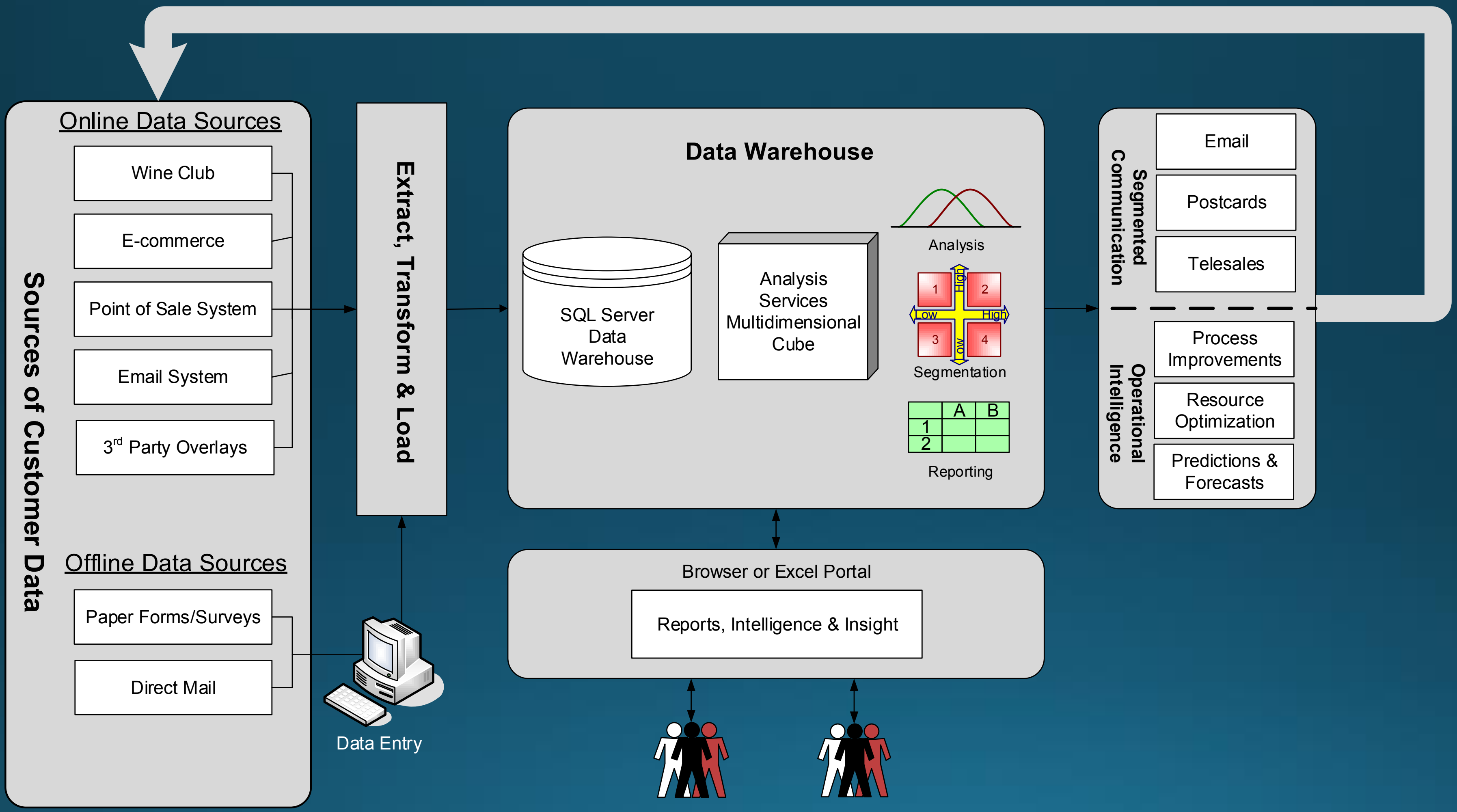
Project Plan

COLLECT
Gather/Clean Data
Finish Data Extract Programming
Check Data Integrity
Develop Data Dictionary

UNDERSTAND
Data Discovery/Descriptive Stats
Seek Clarification from SME's
Develop/Document Methodology
Data Transformation/Modeling

SYNTHESIZE
Evaluate Significance
Rationalize Results
Evaluate Practicality
Present Results

Data Collection Phase



Data Collection Phase

1. Created 2 complete data warehouses using SQL Server 2012 BI Edition
   - All billing addresses are verified and standardized using 3rd-party software (*NetZipCode by The Software Company*)
   - All contact names are parsed and genderized (*NetGender by The Software Company*)
   - Email and phone numbers formats are checked using Regular Expressions
   - Contact records are de-duplicated based on parts of last name, street address and zip code and/or email address.
   - Includes all sales transactions down to the item level
   - Includes order notes, customer notes and delivery status notes since 2012
   - Include all email opens, clicks, bounces, unsubscribes and non-responses since 2008
   - Included 224 census variables joined at the zip code level (*Zip Code USA*)
   - Included 42 purchased demographic variables and last move date (Acxiom)

Data Collection Phase

2. Wrote SQL Queries to create analysis dataset for each winery
    - Included all wine club members that were active (not cancelled) as of 1/1/2010
    - Sales & Email transactional data included through 12/31/2013
    - Summarized sales data by:
        - Wine Club Sales / Non-Wine Club Sales / Total Sales
    - Summarized email data by:
        - Promotion-oriented emails / General information emails / Total emails
    - Created 48 monthly snapshots at customer level (i.e., 1/2010 through 12/2013)
        - Summed lifetime-to-date sales and email variables
        - Captured the prior month's activity
        - Captured all account-level changes and notes for each snapshot
        - Created 2 target cancellation target variables – 3 months and 6 months into the future

Data Collection Phase

3. Tweaked data collection approach after reviewing data
   - Recognized that we don't have near enough data to predict on a monthly snapshot level. Only 2% cancel on a monthly basis. Rare event.
   - Created 1 modeling file for each winery:
     - Snapshot of cancelled club members as they looked the month that they cancelled. Binary target value = 1 (cancelled).
     - Snapshot of non-cancelled club members as they looked on 12/2013. Binary target value = 0 (not cancelled).
4. Standardized summarized sales and count variables
   - Total sales and count measures for non-cancelled members would increase the longer they remained active which would be collinear with Months Since Club Start.
   - Standardized each variable by dividing by the months since first activity. Example: Cumulative Wine Club Sales is standardized by taking Cumulative Wine Club Sales divided by Number of Months Since First Wine Club Sale.

Data Discovery Phase

1. Created SAS Macro to automate plotting of interval variables
   - Histogram → Assess normality
     - For non-normal data, look for optimal categorization of data (e.g., age, miles from winery)
   - Scatterplot across Target → Assess patterns / correlation
     - Investigate variables that are overly correlated with target (collinearity)
   - Boxplots across Target → Assess differences in mean, median and variation
     - Identify large difference in variances indicating potential non-linear relationship. Highlight for further investigation
2. Created SAS Macro to help identify interactions
   - Runs linear regression on each unique combination of interval predictor variables (Y by X) across target variable . Saves parameters and standard errors estimates to macro variables
   - Uses PROC SQL to standardize beta estimate using standard error and then calculates differences across target variable (i.e., large difference between the slope of the bivariate fit across the binary target).
   - Prints sorted report to identify potential interactions to investigate further

Data Discovery Phase

3. Used JMP to visually assess categorical variables across target
    - Where possible, tried to collapse categories down to binary variables
        - Primarily used judgment but also looked at decision tree splits
    - Standardized similar variables across winery
        - Example: Wine Club Tier was decomposed to bottles per shipment, frequency and base-club/special-club indicators. Winery A and Winery B have very different clubs. This was an attempt to generalize predictor variables across wineries.
    - Fixed any missing values through imputation (very few missing values)
    - Removed fields that were junk or highly dimensional
        - Zip Code, CSA, CBSA, PMSA, etc.
    - Removed text notes and shipment delivery data because it was only available for 2 of the 4 years being studied. Will run separate analysis.

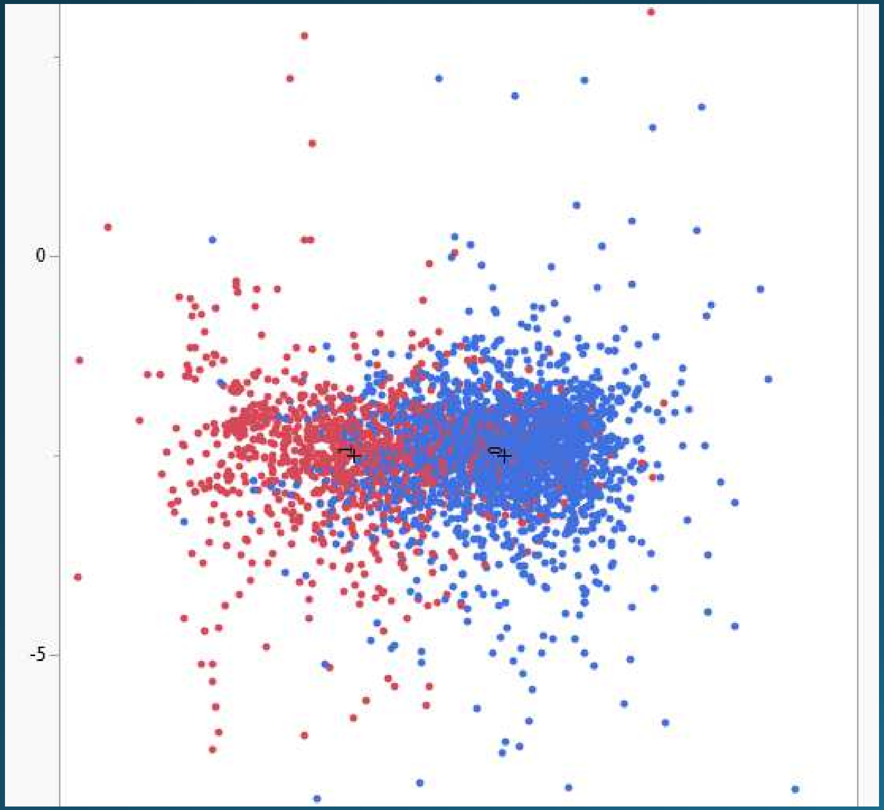
Data Discovery Phase

4. Used Principal Components on Zip-level Census Data
   - Needed a way to create distinct categories from 224 predictor variables
   - Used SAS Enterprise Miner with default settings (correlation matrix). Used SAS node to save 5 principal components back to SAS dataset

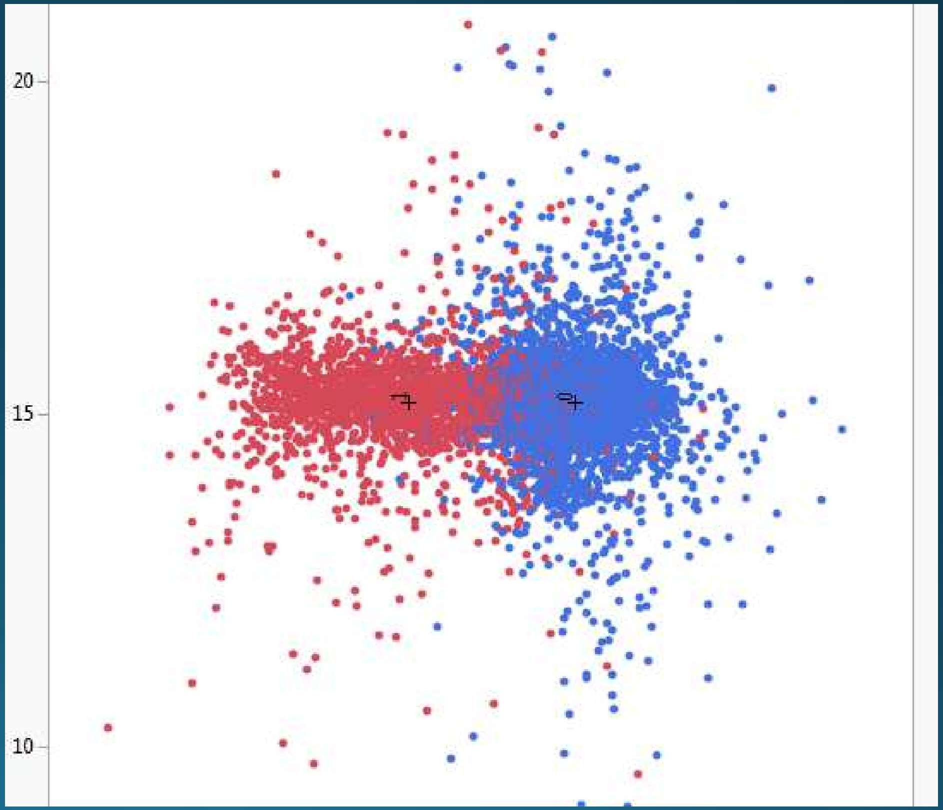
Logistic Modeling Phase

1. Ran Discriminant Analysis to Evaluate Separation

Winery A



Winery B



Logistic Modeling Phase

2. Investigated Potential Quadratics Identified in Data Discovery

Quadratics with Signficant p-values

Winery A	**Winery B**
Avg_Club_ItemPrice*Avg_Club_ItemPrice	Avg_Club_ItemPrice*Avg_Club_ItemPrice
Avg_Club_ItemsPerOrder*Avg_Club_ItemsPerOrder	Avg_Club_ItemsPerOrder*Avg_Club_ItemsPerOrder
Avg_Club_SalesPerOrder*Avg_Club_SalesPerOrder	Avg_Club_SalesPerOrder*Avg_Club_SalesPerOrder
Avg_NonClub_ItemPrice*Avg_NonClub_ItemPrice	Avg_NonClub_ItemPrice*Avg_NonClub_ItemPrice
Avg_NonClub_ItemsPerOrder*Avg_Club_ItemsPerOrder	Avg_NonClub_ItemsPerOrder*Avg_NonClub_ItemsPerOrder
Avg_NonClub_SalesPerOrder*Avg_NonClub_SalesPerOrder	Avg_NonClub_SalesPerOrder*Avg_NonClub_SalesPerOrder
Cumu_ALL_DiscPct*Cumu_ALL_DiscPct	Cumu_ALL_DiscPct*Cumu_ALL_DiscPct
Cumu_Club_DiscPct*Cumu_Club_DiscPct	Cumu_Club_DiscPct*Cumu_Club_DiscPct
Cumu_NonClub_DiscPct*Cumu_NonClub_DiscPct	Cumu_NonClub_DiscPct*Cumu_NonClub_DiscPct
STD_Cuml_All_Clicks*STD_Cuml_All_Clicks	MonthsSinceClubStart*MonthsSinceClubStart
STD_Cuml_ALL_Disc*STD_Cuml_ALL_Disc	MonthsSinceLast_Club*MonthsSinceLast_Club
STD_Cuml_All_DiscountOffers*STD_Cuml_All_DiscountOffers	MonthsSinceLast_NonClub*MonthsSinceLast_NonClub
STD_Cuml_ALL_Items*STD_Cuml_ALL_Items	STD_Cuml_All_Clicks*STD_Cuml_All_Clicks
STD_Cuml_ALL_Net*STD_Cuml_ALL_Net	STD_Cuml_ALL_Disc*STD_Cuml_ALL_Disc
STD_Cuml_All_NoResponses*STD_Cuml_All_NoResponses	STD_Cuml_All_DiscountOffers*STD_Cuml_All_DiscountOffers
STD_Cuml_All_Opens*STD_Cuml_All_Opens	STD_Cuml_ALL_Items*STD_Cuml_ALL_Items
STD_Cuml_ALL_Orders*STD_Cuml_ALL_Orders	STD_Cuml_ALL_Net*STD_Cuml_ALL_Net
STD_Cuml_All_ShippingOffers*STD_Cuml_All_ShippingOffers	STD_Cuml_All_Opens*STD_Cuml_All_Opens
STD_Cuml_Club_Disc*STD_Cuml_Club_Disc	STD_Cuml_ALL_Orders*STD_Cuml_ALL_Orders
STD_Cuml_Club_Items*STD_Cuml_Club_Items	STD_Cuml_All_ShippingOffers*STD_Cuml_All_ShippingOffers
STD_Cuml_Club_Net*STD_Cuml_Club_Net	STD_Cuml_Club_Disc*STD_Cuml_Club_Disc
STD_Cuml_Club_Orders*STD_Cuml_Club_Orders	STD_Cuml_Club_Items*STD_Cuml_Club_Items
STD_Cuml_NonClub_Disc*STD_Cuml_NonClub_Disc	STD_Cuml_Club_Net*STD_Cuml_Club_Net
STD_Cuml_NonClub_Items*STD_Cuml_NonClub_Items	STD_Cuml_Club_Orders*STD_Cuml_Club_Orders
STD_Cuml_NonClub_Net*STD_Cuml_NonClub_Net	STD_Cuml_NonClub_Disc*STD_Cuml_NonClub_Disc
STD_Cuml_NonClub_Orders*STD_Cuml_NonClub_Orders	STD_Cuml_NonClub_Items*STD_Cuml_NonClub_Items
	STD_Cuml_NonClub_Net*STD_Cuml_NonClub_Net
	STD_Cuml_NonClub_Orders*STD_Cuml_NonClub_Orders

← Highlighted variables are common between wineries

Logistic Modeling Phase

3. Investigated Potential Interactions Identified in Data Discovery

Interactions with Signficant p-values

Winery A

MonthsSinceFirst_EmailALL*Avg_Club_ItemPrice
MonthsSinceFirst_ALL*STD_Cuml_All_DiscountOffers
MonthsSinceFirst_Club*STD_Cuml_All_DiscountOffers
MonthsSinceFirst_EmailALL*STD_Cuml_All_DiscountOffers
MonthsSinceLast_NonClub*Avg_Club_ItemPrice
MonthsSinceLast_NonClub*STD_Cuml_All_DiscountOffers
Cumu_NonClub_DiscPct*Avg_Club_ItemPrice
Cumu_NonClub_DiscPct*STD_Cuml_All_DiscountOffers
MonthsSinceLast_ALL*Avg_Club_ItemPrice
Avg_NonClub_SalesPerOrder*STD_Cuml_All_DiscountOffers
STD_Cuml_All_Reminders*MonthsSinceLast_NonClub
Avg_NonClub_ItemPrice*STD_Cuml_All_DiscountOffers
STD_Cuml_All_DiscountOffers*Avg_NonClub_ItemsPerOrder
STD_Cuml_Club_Orders*STD_Cuml_All_ShippingOffers
MonthsSinceLast_Club*MonthsSinceLast_EmailALL
MonthsSinceLast_NonClub*STD_Cuml_All_Opens
Avg_NonClub_SalesPerOrder*STD_Cuml_All_Reminders
MonthsSinceLast_ALL*STD_Cuml_All_Reminders

Winery B

MonthsSinceFirst_Club*STD_Cuml_Club_Disc
STD_Cuml_All_DiscountOffers*STD_Cuml_All_ShippingOffers
STD_Cuml_Club_Net*STD_Cuml_Club_Disc
STD_Cuml_NonClub_Net*STD_Cuml_NonClub_Disc
STD_Cuml_Club_Items*STD_Cuml_Club_Disc
MonthsSinceFirst_ALL*STD_Cuml_All_Reminders
STD_Cuml_Club_Net*STD_Cuml_All_DiscountOffers
STD_Cuml_All_DiscountOffers*STD_Cuml_Club_Items
STD_Cuml_Club_Disc*MonthsSinceClubStart
STD_Cuml_All_Reminders*MonthsSinceLast_NonClub
STD_Cuml_NonClub_Disc*STD_Cuml_ALL_Items
Avg_NonClub_ItemsPerOrder*STD_Cuml_NonClub_Disc
STD_Cuml_All_Reminders*MonthsSinceLast_ALL
STD_Cuml_All_ShippingOffers*STD_Cuml_Club_Orders
STD_Cuml_NonClub_Items*Cumu_ALL_DiscPct
Cumu_ALL_DiscPct*Avg_NonClub_SalesPerOrder
STD_Cuml_All_DiscountOffers*STD_Cuml_ALL_Net
STD_Cuml_ALL_Items*Cumu_ALL_DiscPct
STD_Cuml_ALL_Net*Cumu_ALL_DiscPct
STD_Cuml_NonClub_Disc*STD_Cuml_All_DiscountOffers
MonthsSinceLast_NonClub*STD_Cuml_All_ShippingOffers
Avg_Club_SalesPerOrder*STD_Cuml_All_DiscountOffers

← Highlighted variables are common between wineries

Logistic Modeling Phase

4. Used JMP to fit Logistic Regression Models

- Created Validation / Training columns (60% training / 40% Validation)
- Included Main Effects and selected Quadratics & Interactions
- Used Forward P-value, Forward BIC, Mixed P-value and Max Validation R^2 model selection techniques.
    - The Backward method would not converge (Step-halving limit)
    - Manually removed non-significant variables from training models
- Looked at validation misclassification rate as well as true positive and true negative rates to assess fit.
    - Considered the tradeoff between a larger model and improvements in misclassification

Logistic Modeling Phase

5. Selected model Training data fit – Winery A

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Last_ClubOrder_GT3months[1]	3.829101	0.391809	95.51	<.0001*
MonthsSinceClubStart	0.102001	0.013562	56.56	<.0001*
Avg_Club_ItemPrice	-0.08521	0.011784	52.28	<.0001*
MonthsSinceFirst_EmailALL_GT24[1]	-1.03459	0.152276	46.16	<.0001*
(MonthsSinceClubStart-33.8235)*(MonthsSinceClubStart-33.8235)	-0.00095	0.000149	40.24	<.0001*
MonthsSinceFirst_ALL	-0.13887	0.021955	40.01	<.0001*
Recvd_Offer_Last1Months[1]	-0.85057	0.134676	39.89	<.0001*
MonthsSinceLast_Club	-0.81555	0.13226	38.02	<.0001*
(MonthsSinceLast_Club-3.15599)*(MonthsSinceLast_Club-3.15599)	0.040408	0.006625	37.20	<.0001*
(MonthsSinceFirst_ALL-23.5952)*(STD_Cuml_All_DiscountOffers-1.44463)	0.065473	0.011317	33.47	<.0001*
(STD_Cuml_All_ShippingOffers-1.0146)*(STD_Cuml_Club_Orders-0.35357	3.994778	0.775057	26.57	<.0001*
STD_Cuml_All_NoResponses	0.464669	0.097572	22.68	<.0001*
Avg_Club_ItemPrice_GT40[1]	0.796706	0.178752	19.87	<.0001*
MonthsSinceFirst_ALL_GT38[1]	-1.05672	0.237876	19.73	<.0001*
STD_Cuml_All_DiscountOffers	0.975295	0.230101	17.97	<.0001*
(Avg_Club_ItemsPerOrder-3.81659)*(Avg_Club_ItemsPerOrder-3.81659)	0.03784	0.009379	16.28	<.0001*
IsClubOnHold[1]	-2.45727	0.614917	15.97	<.0001*
Avg_Club_ItemsPerOrder	-0.31222	0.081077	14.83	0.0001*
STD_Cuml_All_ShippingOffers	-0.57826	0.162282	12.70	0.0004*
STD_Cuml_Club_Orders	2.855337	0.809983	12.43	0.0004*
Intercept	3.53299	1.283081	7.58	0.0059*
Ever_NoResponse[1]	0.399119	0.172197	5.37	0.0205*

For log odds of 1/0

Logistic Modeling Phase

5. Selected model Training data fit – Winery B

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
STD_Cuml_All_Reminders	-9.631328	0.9765637	97.27	<.0001*
STD_Cuml_All_Opens	2.14672876	0.2205993	94.70	<.0001*
Intercept	25.9402133	2.8650896	81.97	<.0001*
STD_Cuml_All_NoResponses	1.90243785	0.2171145	76.78	<.0001*
STD_Cuml_Club_Disc	0.67864079	0.0798893	72.16	<.0001*
Avg_Club_ItemPrice	-0.6143825	0.0785344	61.20	<.0001*
MilesFromWinery_GT100[1]	0.77706775	0.1073412	52.41	<.0001*
Cumu_Club_DiscPct	0.28163626	0.0390569	52.00	<.0001*
Avg_Club_ItemsPerOrder	-4.2980506	0.6166689	48.58	<.0001*
Avg_Club_SalesPerOrder	0.11778271	0.0176163	44.70	<.0001*
MonthsSinceFirst_Club	-0.196084	0.0331075	35.08	<.0001*
STD_Cuml_Club_Items	-1.647882	0.2812174	34.34	<.0001*
Had_ClubOrder_Last3Months[1]	-0.9052184	0.1591926	32.33	<.0001*
Ever_NoResponse[1]	1.03792683	0.192511	29.07	<.0001*
Cumu_NonClub_DiscPct_GT0[1]	-0.4015944	0.075673	28.16	<.0001*
Recvd_Offer_Last1Months[1]	-0.8928491	0.1700803	27.56	<.0001*
STD_Cuml_All_ShippingOffers	-3.3682415	0.6669801	25.50	<.0001*
MonthsSinceLast_Club	-0.1707746	0.0344619	24.56	<.0001*
(STD_Cuml_Club_Items-1.4453)*(STD_Cuml_Club_Items-1.4453)	0.35079917	0.0713268	24.19	<.0001*
MonthsSinceFirst_Club_GT38[1]	0.61846875	0.1538522	16.16	<.0001*
(STD_Cuml_All_ShippingOffers-0.47225)*(STD_Cuml_All_ShippingOffers-0.47225	3.34335646	0.944236	12.54	0.0004*
STD_Cuml_Club_Items_GT1[1]	-0.3693969	0.1101996	11.24	0.0008*
MonthsSinceFirst_ALL	-0.0837471	0.0317398	6.96	0.0083*
(Avg_Club_ItemPrice-32.3626)*(Avg_Club_ItemPrice-32.3626)	0.01071394	0.0041196	6.76	0.0093*
MonthsSinceFirst_NonClub	-0.0154411	0.0065316	5.59	0.0181*
STD_Cuml_Club_Net_GT40[1]	0.23974925	0.1184606	4.10	0.0430*

For log odds of 1/0

Logistic Modeling Phase

5. Selected model Validation data fit – Winery A

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Last_ClubOrder_GT3months[1]	2.96900218	0.3763348	62.24	<.0001*
Avg_Club_ItemPrice	-0.0686503	0.0120983	32.20	<.0001*
MonthsSinceClubStart	0.09593869	0.0177301	29.28	<.0001*
(Avg_Club_ItemsPerOrder-3.81251)*(Avg_Club_ItemsPerOrder-3.81251)	0.06185406	0.0120689	26.27	<.0001*
MonthsSinceFirst_ALL	-0.1415965	0.0282999	25.03	<.0001*
(MonthsSinceLast_Club-3.28021)*(MonthsSinceLast_Club-3.28021)	0.01847291	0.0037135	24.75	<.0001*
(MonthsSinceClubStart-33.7635)*(MonthsSinceClubStart-33.7635)	-0.0010675	0.0002179	24.00	<.0001*
MonthsSinceFirst_EmailALL_GT24[1]	-0.7756738	0.1710178	20.57	<.0001*
Avg_Club_ItemsPerOrder	-0.4897605	0.1099541	19.84	<.0001*
MonthsSinceLast_Club	-0.4616904	0.1112078	17.24	<.0001*
(MonthsSinceFirst_ALL-23.6116)*(STD_Cuml_All_DiscountOffers-1.44769)	0.05181686	0.0130337	15.81	<.0001*
Avg_Club_ItemPrice_GT40[1]	0.6618244	0.1815623	13.29	0.0003*
(STD_Cuml_All_ShippingOffers-1.01192)*(STD_Cuml_Club_Orders-0.35117)	2.9840402	0.8742246	11.65	0.0006*
Recvd_Offer_Last1Months[1]	-0.5237691	0.1597523	10.75	0.0010*
MonthsSinceFirst_ALL_GT38[1]	-0.9034748	0.2762351	10.70	0.0011*
STD_Cuml_Club_Orders	2.88895828	0.9400641	9.44	0.0021*
IsClubOnHold[1]	-1.6013969	0.535068	8.96	0.0028*
Intercept	3.71456228	1.3524429	7.54	0.0060*
STD_Cuml_All_ShippingOffers	-0.4955241	0.1860753	7.09	0.0077*
Ever_NoResponse[1]	0.48089413	0.1991398	5.83	0.0157*
STD_Cuml_All_NoResponses	0.24589028	0.112456	4.78	0.0288*
STD_Cuml_All_DiscountOffers	0.50420081	0.2592509	3.78	0.0518

For log odds of 1/0

Logistic Modeling Phase

5. Selected model Validation data fit – Winery B

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
STD_Cuml_Club_Disc	0.90733452	0.1065139	72.56	<.0001*
STD_Cuml_All_Reminders	-10.58711	1.3784664	58.99	<.0001*
Intercept	26.6171937	3.7295048	50.94	<.0001*
MonthsSinceFirst_Club	-0.2829162	0.0399271	50.21	<.0001*
STD_Cuml_All_NoResponses	1.86160046	0.2663367	48.86	<.0001*
STD_Cuml_Club_Items	-2.6214609	0.3770172	48.35	<.0001*
(STD_Cuml_Club_Items-1.46596)*(STD_Cuml_Club_Items-1.46596)	0.63122086	0.0976554	41.78	<.0001*
STD_Cuml_All_Opens	1.67058941	0.2772699	36.30	<.0001*
Cumu_Club_DiscPct	0.30029103	0.0518868	33.49	<.0001*
MilesFromWinery_GT100[1]	0.72238873	0.1316393	30.11	<.0001*
Avg_Club_ItemPrice	-0.5547663	0.1036826	28.63	<.0001*
MonthsSinceFirst_Club_GT38[1]	1.03022219	0.1985339	26.93	<.0001*
MonthsSinceLast_Club	-0.2115964	0.049082	18.59	<.0001*
Avg_Club_ItemsPerOrder	-3.3736742	0.8016769	17.71	<.0001*
Ever_NoResponse[1]	1.04781717	0.2516152	17.34	<.0001*
Avg_Club_SalesPerOrder	0.09053464	0.0235865	14.73	0.0001*
Recvd_Offer_Last1Months[1]	-0.8426719	0.2230036	14.28	0.0002*
Had_ClubOrder_Last3Months[1]	-0.7361182	0.2057208	12.80	0.0003*
STD_Cuml_All_ShippingOffers	-2.7983387	0.8684329	10.38	0.0013*
MonthsSinceFirst_NonClub	-0.0225311	0.0083197	7.33	0.0068*
STD_Cuml_Club_Net_GT40[1]	0.35806063	0.1499867	5.70	0.0170*
Cumu_NonClub_DiscPct_GT0[1]	-0.2090313	0.1001151	4.36	0.0368*
MonthsSinceFirst_ALL	-0.0638511	0.0349325	3.34	0.0676
(Avg_Club_ItemPrice-32.365)*(Avg_Club_ItemPrice-32.365)	0.00834604	0.0047157	3.13	0.0768
STD_Cuml_Club_Items_GT1[1]	-0.2394282	0.143602	2.78	0.0955
(STD_Cuml_All_ShippingOffers-0.47199)*(STD_Cuml_All_ShippingOffers-0.47199	1.42217907	1.1780317	1.46	0.2273

For log odds of 1/0

Logistic Modeling Phase

6. Are parameters intuitive? Winery A

When the following INCREASES:	Attrition Risk:	Immediately Intuitive?	Potential Explanation:
Last wine club order was > 3 months ago	↑	Yes	No recent club orders likely means credit card declined
Average price per item in club shipment	↓	No	Primary club tier was discontinued and replaced by higher priced in 2012
Number of months since club start date	↑	Yes	Quadratic effect. Increases risk until 33.76 months and then lowers risk
Number of months since first purchase	↓	No	First purchase likely before club membership began - loyal ambassadors
First email sent > 24 months ago	↓	Yes	Indicates we've had working email address for quite some time. Lowers risk.
Average items per club shipment	↓	No	Quadratic effect. Decreases risk until 3.81 items and then increases risk
Number of months since last club shipment	↓	No	Quadratic effect. Decreases attrition risk until 3.28 months and then increases risk
Average price per item in club shipment > $40	↑	Yes	Very high priced ($90/bottle) Vintner Select case club was discontinued
Received email promotional offer last month	↓	Yes	Members receiving recent email promotion
Number of months since first purchase > 38	↓	Yes	First purchase likely before club membership began - loyal ambassadors
Standardized cumulative club orders	↑	Yes	More club orders equates to more time in club and increase risk
Club is on hold	↓	Yes	Clubs that are on hold are just skipping a few shipments. Will be reactivated.
Standardized cumlative email shipping discount offers	↓	Yes	Members that receive discounted shipping offers have lower risk. Deal seekers.
At least one email was NOT opened	↑	Yes	Members that don't open emails are less engaged and higher risk
Standardized cumulative email non responses	↑	Yes	Members that don't open emails are less engaged and higher risk
Standardized cumulative email product discount offers	↑	No	Members that receive product discount offers may feel overcharged for club shipment

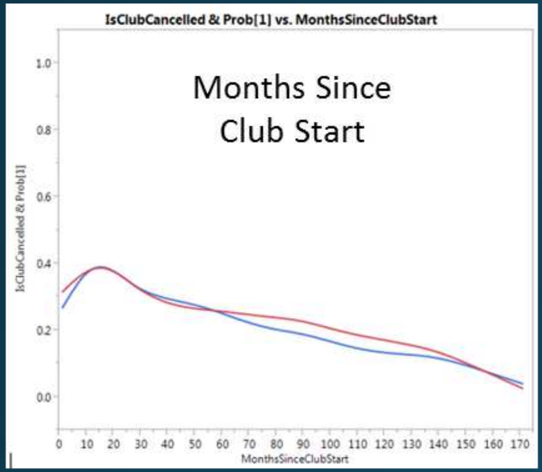
Logistic Modeling Phase

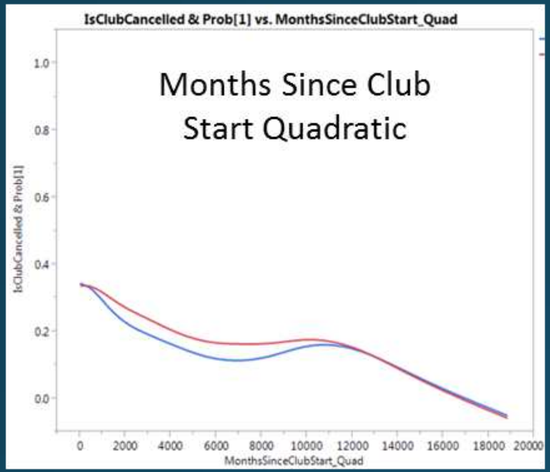
6. Are parameters intuitive? Winery B

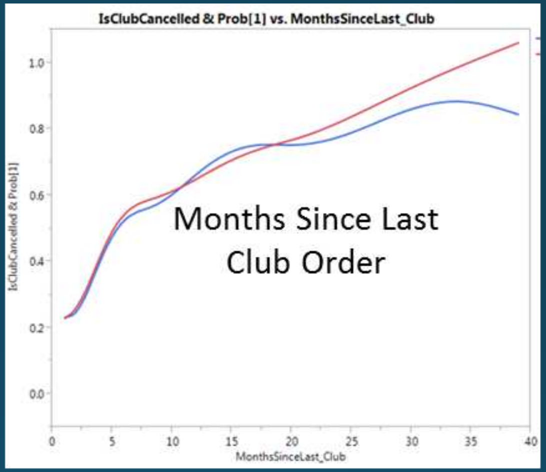
When the following INCREASES:	Attrition Risk:	Immediately Intuitive?	Potential Explanation:
Standardized cumulative club discount	↑	Yes	Discounts have slowly been taken away. Price senstitive members are leaving.
Standardized cumulative email reminders	↓	No	Email offers have been greatly reduced in lieu of telesales. Very few email reminders.
Number of months since first club purchase	↓	No	Members that have been transacting for a long time are lower risk (loyal)
Standardized cumulative email no responses	↑	Yes	Members that don't open emails are less engaged and higher risk
Standardized cumulative club items	↓	No	Quadratic effect.Decreases risk until 1.47 items and then increases risk
Standardized cumulative email opens	↑	No	Emails are very winery news & event oriented. May be alienating distant customers.
Cumulative average club discount percent	↑	No	Members with a high average discount % are higher risk. May be price sensitive
Miles from winery > 100	↑	Yes	Strong local following due to winery events. Less strong for distant members
Average club item price	↓	Yes	Quadratic effect. Decreases risk until $32.37 and then increases risk
Number of months since first club purchase > 38	↑	No	Length of time since membership started increases risk
Months since last club order	↓	No	Could be impact of club option to consolidate shipments (ship fewer times per year)
Average club items per order	↓	Yes	Large base of loyal club members participating at minimum level
At least one email was NOT opened	↑	Yes	Members that don't open email are less engaged and higher risk
Average club sales per order	↑	Yes	Average cost of club has been rising leading to increased risk
Received email offer in last month	↓	Yes	Members receiving recent email promotion show lower risk
Had a club order in past 3 months	↓	Yes	Members with high recency are lower risk
Standardized cumulative email shipping offers	↓	No	Quadratic effect. Decreases risk until .47 shipping offers and then increases risk
Number of months since first non-club purchase	↓	No	First purchase likely before club began - loyal ambassadors
Standardized cumulative net sales > $40	↑	Yes	Members that average $40 per month are higher risk. May want discount for large purchase
Average cumulative non-club discount percent > 0%	↓	Yes	Members that utilize their club discount for a la carte wines are lower risk
Number of months since first order of any type	↓	No	First purchase likely before club began - loyal ambassadors
Standardized cumulative club items > 1	↓	Yes	Members that average 1+ bottle of club wine per month are lower risk

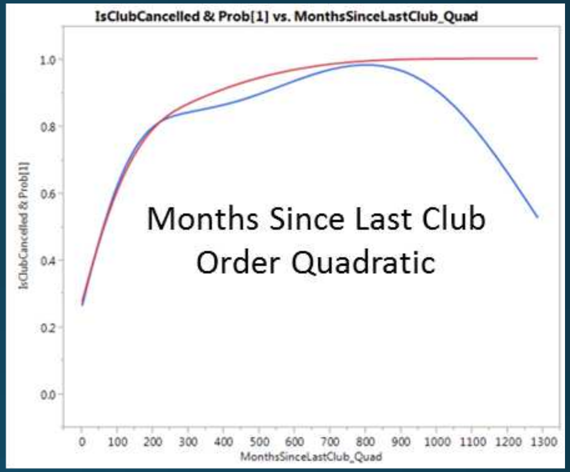
Logistic Modeling Phase

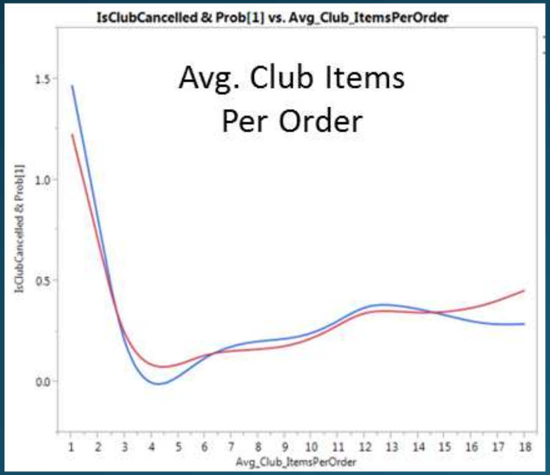
7. Assess Validity of Model – Winery A Marginal Model Plots #1

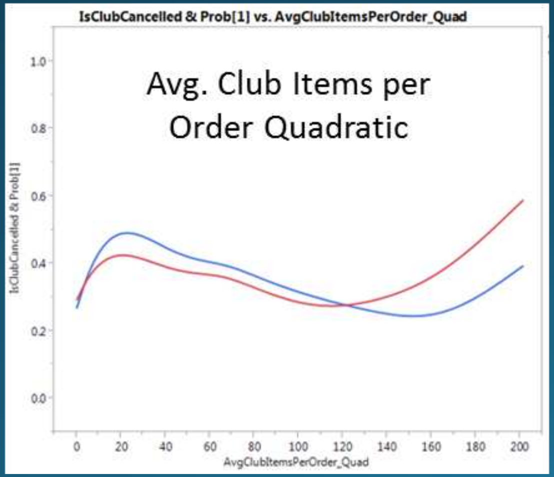


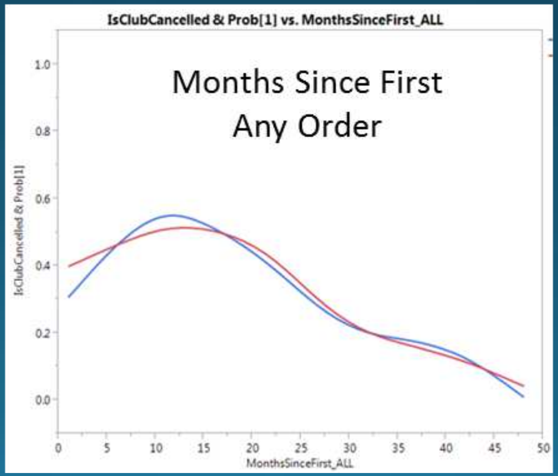


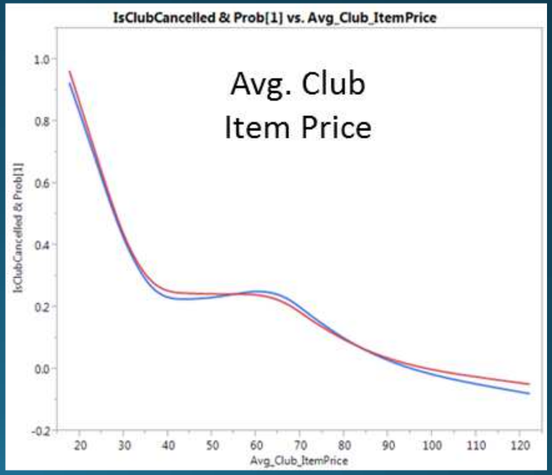






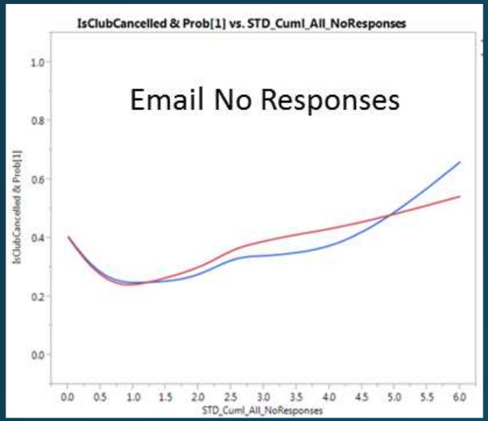


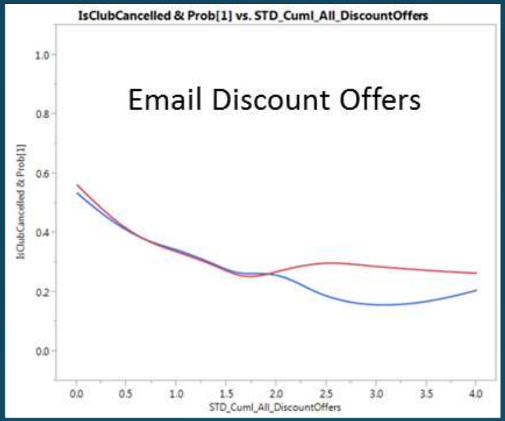


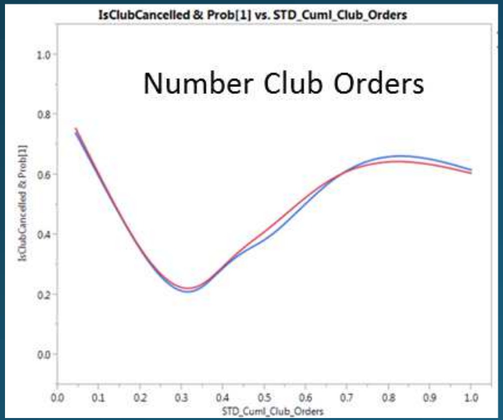


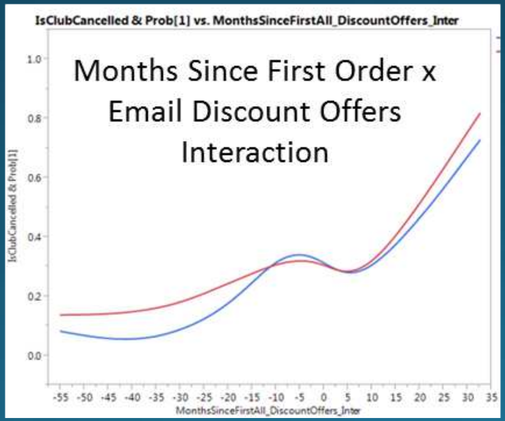
Logistic Modeling Phase

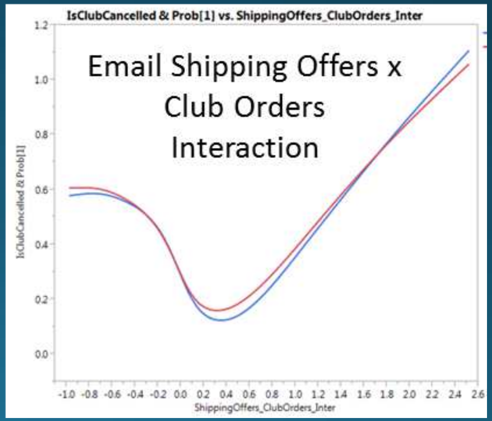
7. Assess Validity of Model – <u>Winery A</u> Marginal Model Plots #2





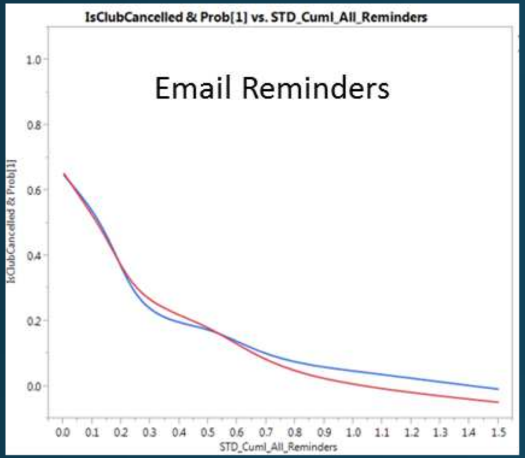


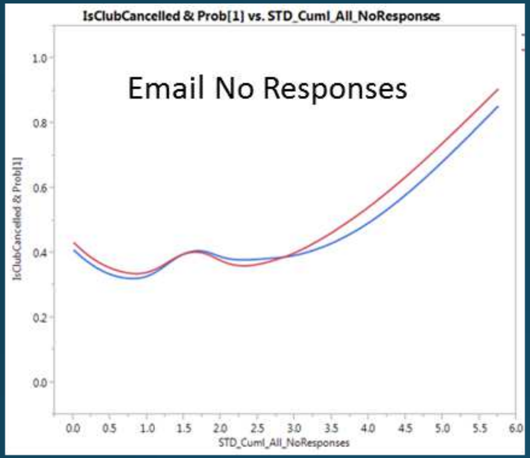


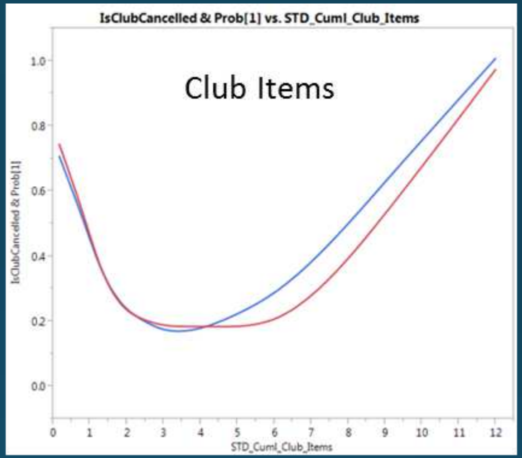


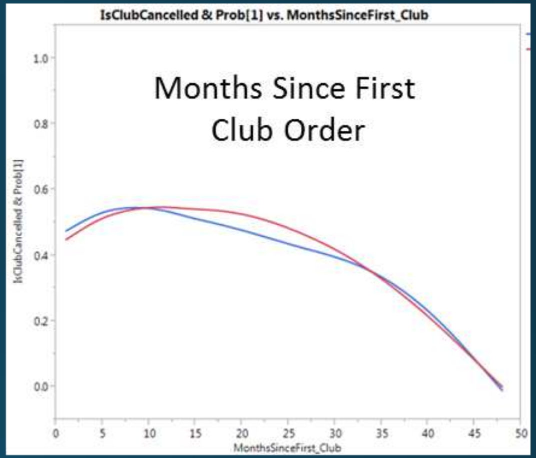
Logistic Modeling Phase

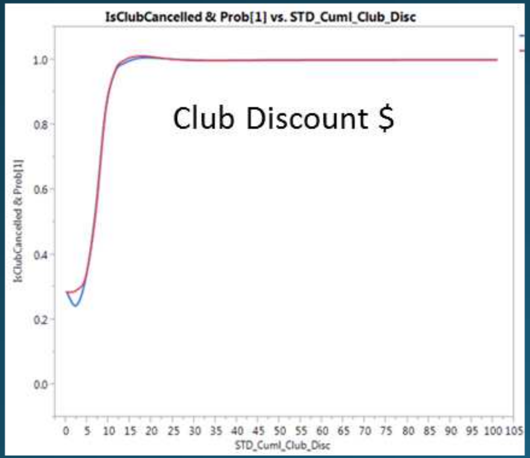
7. Assess Validity of Model – <u>Winery B</u> Marginal Model Plots #1

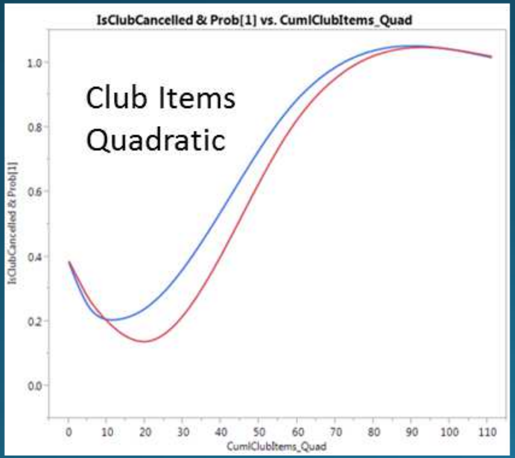


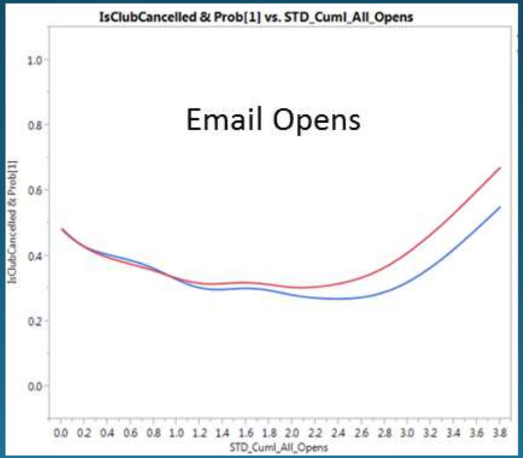


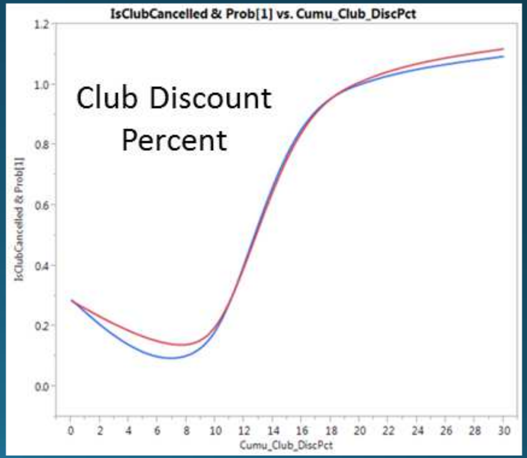






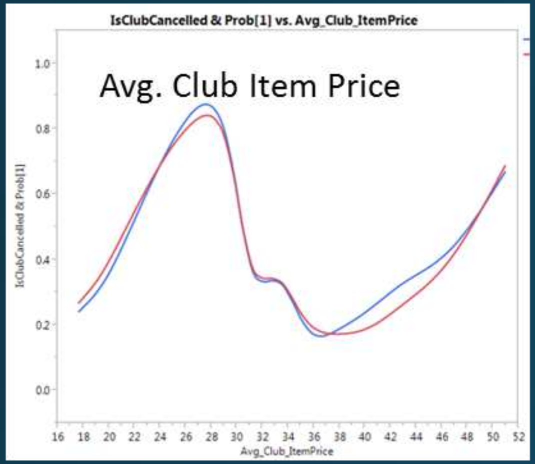


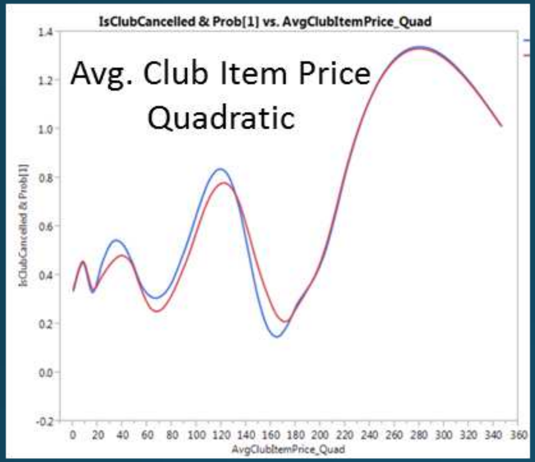


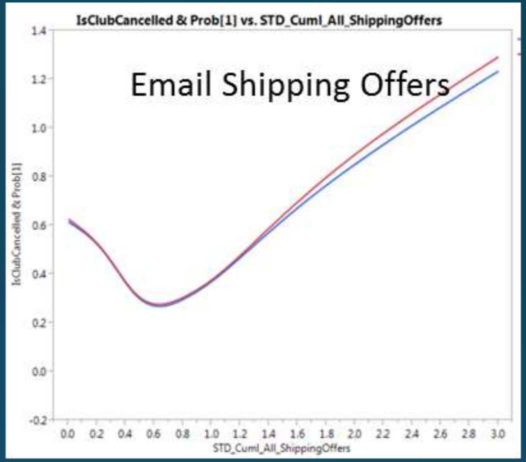


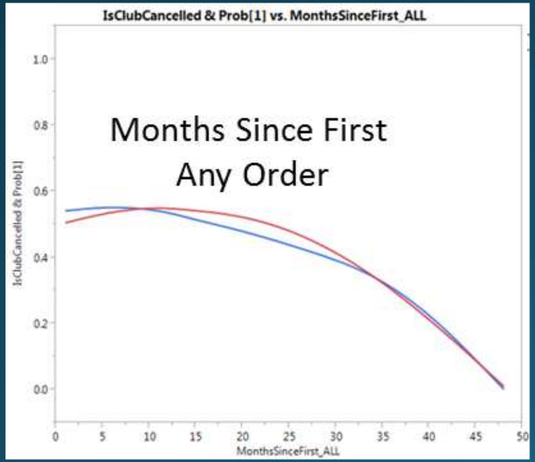
Logistic Modeling Phase

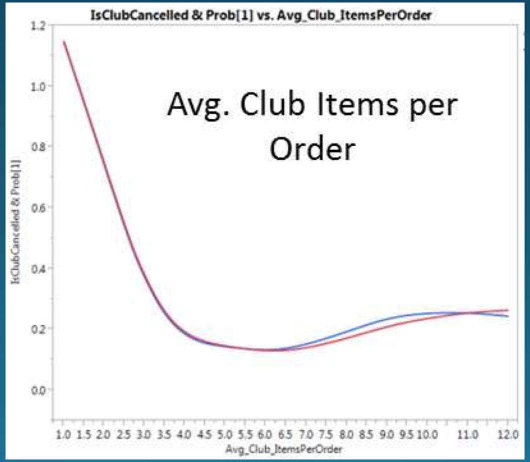
7. Assess Validity of Model – Winery B Marginal Model Plots #2

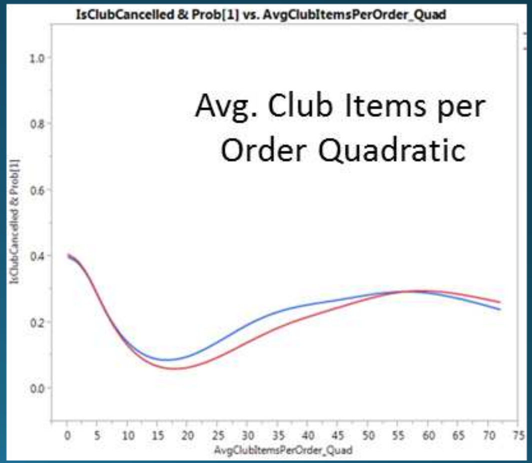








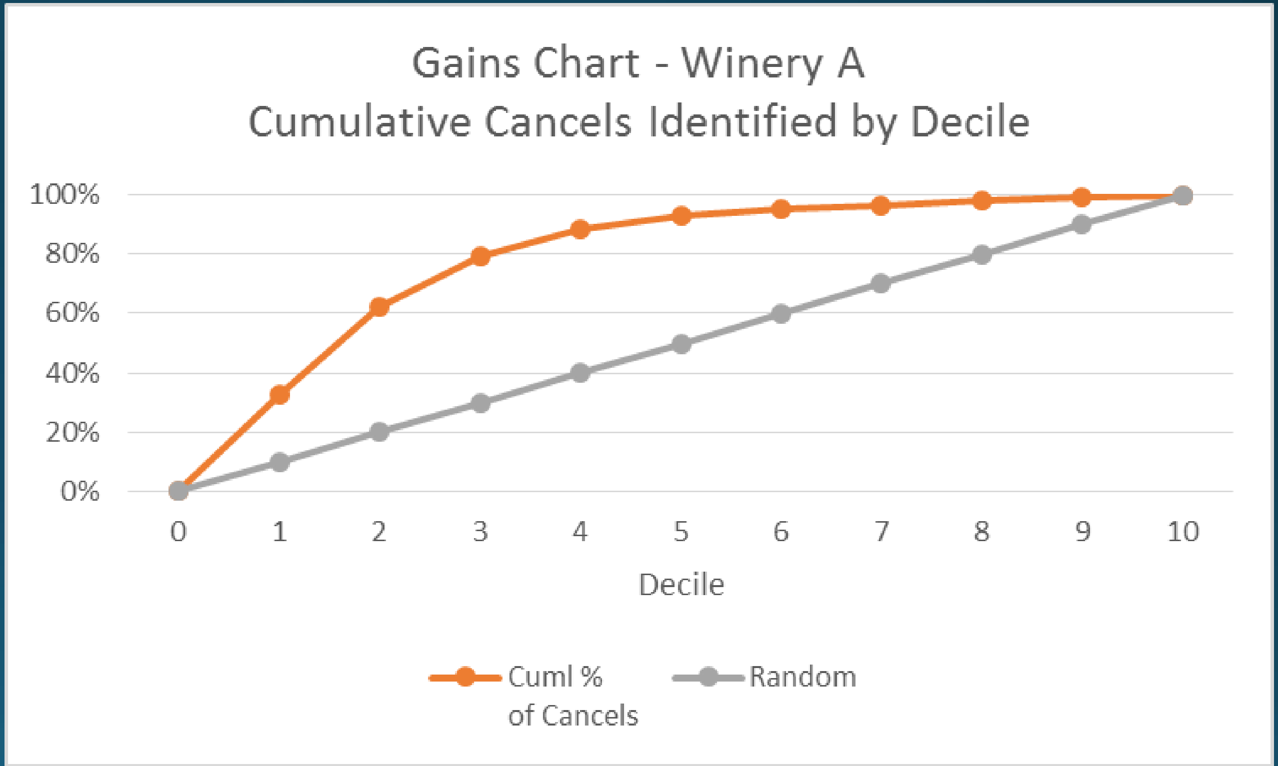




Logistic Modeling Phase

8. Validation model Gains Chart – Winery A

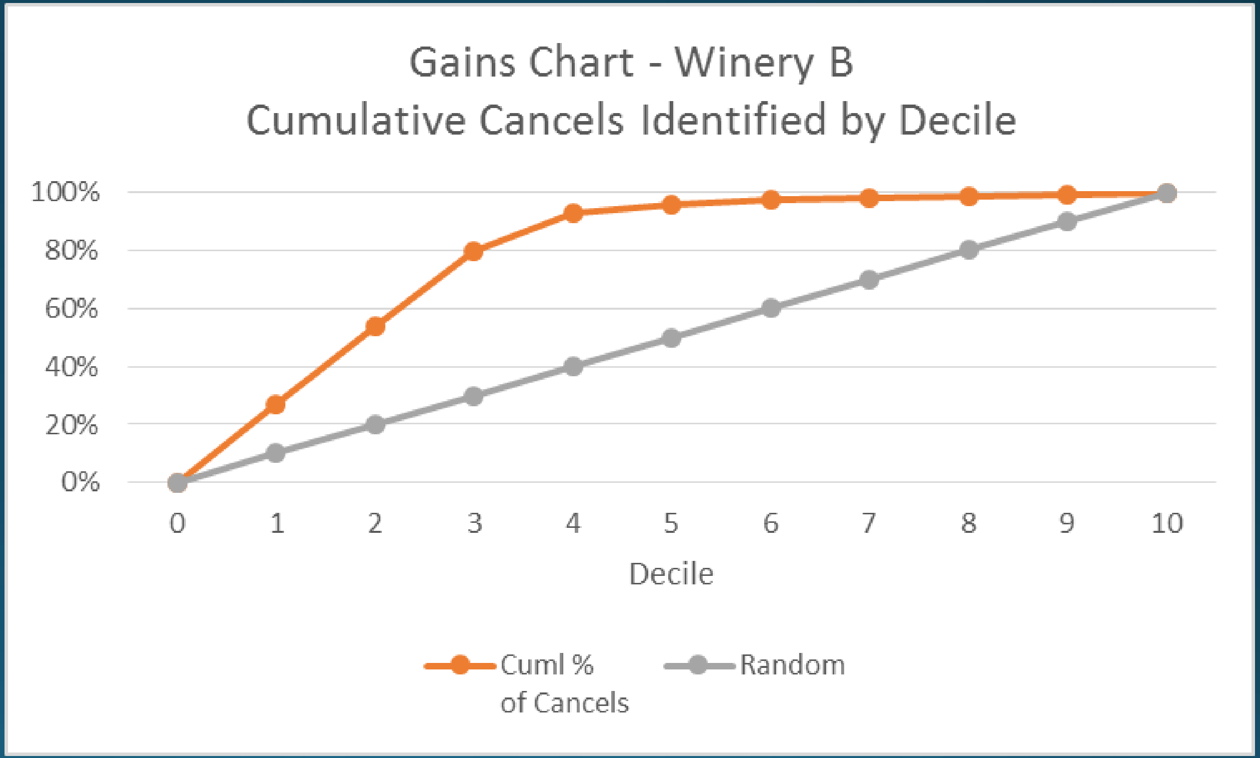
Decile	Cancels	Cuml % of Cancels
1	133	32.4%
2	123	62.4%
3	68	79.0%
4	38	88.3%
5	20	93.2%
6	8	95.1%
7	5	96.3%
8	6	97.8%
9	6	99.3%
10	3	100.0%



Logistic Modeling Phase

8. Validation model Gains Chart – <u>Winery B</u>

Decile	Cancels	Cuml % of Cancels
1	240	26.8%
2	241	53.7%
3	230	79.4%
4	118	92.6%
5	30	96.0%
6	14	97.5%
7	6	98.2%
8	5	98.8%
9	5	99.3%
10	6	100.0%



Logistic Modeling Phase

9. Cutoff Analysis – Validation data

Winery A

Posterior Probability = 0.31

Cutoff: 0.50	Predicted	
Actual	Active	Cancelled
Active	885	54
Cancelled	101	309

Misclass Rate: 11.49%
True Negative Rate: 94.25%
True Positive Rate: 75.37%

Cutoff: 0.45	Predicted	
Actual	Active	Cancelled
Active	866	73
Cancelled	89	321

Misclass Rate: 12.01%
True Negative Rate: 92.23%
True Positive Rate: 78.29%

Cutoff: 0.40	Predicted	
Actual	Active	Cancelled
Active	853	86
Cancelled	82	328

Misclass Rate: 12.45%
True Negative Rate: 90.84%
True Positive Rate: 80.00%

Cutoff: 0.35	Predicted	
Actual	Active	Cancelled
Active	835	104
Cancelled	74	336

Misclass Rate: 13.19%
True Negative Rate: 88.92%
True Positive Rate: 81.95%

Cutoff: 0.30	Predicted	
Actual	Active	Cancelled
Active	802	137
Cancel	60	350

Misclass Rate: 14.60%
True Negative Rate: 85.41%
True Positive Rate: 85.37%

Winery B

Posterior Probability = 0.37

Cutoff: 0.50	Predicted	
Actual	Active	Cancelled
Active	1451	62
Cancelled	115	780

Misclass Rate: 7.35%
True Negative Rate: 95.90%
True Positive Rate: 87.15%

Cutoff: 0.45	Predicted	
Actual	Active	Cancelled
Active	1442	71
Cancelled	106	789

Misclass Rate: 7.35%
True Negative Rate: 95.31%
True Positive Rate: 88.16%

Cutoff: 0.40	Predicted	
Actual	Active	Cancelled
Active	1428	85
Cancelled	92	803

Misclass Rate: 7.35%
True Negative Rate: 94.38%
True Positive Rate: 89.72%

Cutoff: 0.35	Predicted	
Actual	Active	Cancelled
Active	1415	98
Cancelled	84	811

Misclass Rate: 7.56%
True Negative Rate: 93.52%
True Positive Rate: 90.61%

Cutoff: 0.30	Predicted	
Actual	Active	Cancelled
Active	1392	121
Cancelled	72	823

Misclass Rate: 8.01%
True Negative Rate: 92.00%
True Positive Rate: 91.96%

Logistic Modeling Phase

10. Key Learnings

- There were far fewer predictors in common between wineries than I would have anticipated. It appears that underlying club structure and winery-specific processes have a great deal of influence on attrition.
- Wine club members tend to be a very homogeneous group. None of the purchased demographic variables ended up in either model.
- It's critical to have some subject matter experts that have been around awhile. Managerial decisions made in the past can make the interpretation of parameters difficult without context (e.g., clubs being discontinued).
- The customer's geographic location doesn't have much impact. I used Principal Components on 200+ zip-level variables, census divisions & regions, MSA's and Miles From Winery. Only Winery B showed a significant effect for nearby customers (they have many more winery events).

Supplementary Analysis

Supplementary Analysis Overview

1. Would like to know both "if" and "when" a customer will cancel

   - Used Survival Analysis to predicted time-to-event
     - Much of this analysis based on the book *Survival Analysis Using SAS: A Practical Guide, Second Edition* by Paul D. Allison (SAS Press, 2010).
     - Additional insight was gained from SUGI paper #114-27 entitled *Predicting Customer Churn in the Telecommunications Industry – An Application of Survival Analysis Modeling Using SAS* by Junxiang Lu, PHD.
     - Survival Analysis was not covered in any detail in the MS Analytics program. The goal of this analysis is to better understand the method – not produce an optimum model.

2. Would like compare "traditional" logistic modeling to SEM
   - What does the "best" logistic model look like in SAS Enterprise Miner?
   - Similarities & differences from JMP model

Survival Analysis

Survival Analysis

1. Modeling Methodology & Process

- Target variable was MonthsSinceClubStart. Censoring variable was IsClubCancelled (1=Yes, 0=No)
  - Accounts that were still active at end of study were right-censored
  - Unlike logistic, I limited this study to customers that are within 5 years of club start date. I found that too many outliers result in very poor survival estimates.
- Started with the same main effects, quadratics and interactions discovered previously. Removed any effects that could act as a proxy for the target.
- Used semi-parametric stepwise PROC PHREG to decrease the number of effects.
  - Manually removed any remaining terms with p-value > 0.05
  - Evaluated shape of survival distribution
  - Evaluated model significance and goodness of fit

Survival Analysis

1. Modeling Methodology & Process (continued)

- Used JMP to evaluate shape of Log(MonthsSinceClubStart)
- Used parametric PROC LIFEREG to predict survival probabilities
  - Generating predicted event times is cumbersome with PHREG and relatively easy with LIFEREG. However, LIFEREG doesn't handle time-dependent covariates which may be a weakness in my methodology.
  - Built models using different distributions and observed AIC. Selected Weibull.
  - Used Paul Allison "Predict" Macro to calculate survival rates for 6, 12, 18, 24, 30 & 36 month periods.
  - Calculated attrition rate at each period as 1 minus Survival probability
- Validated model with 40% holdout sample
  - Calculated misclassification rate for the period within 24 months of start date
  - Calculated Gains Chart reflecting cumulative cancels up to specified periods

Survival Analysis

2. Semi-parametric model fit using PHREG – Winery A

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
MonthsSinceLast_Club	1	0.05865	0.01115	27.6824	<.0001	1.060
MonthsSinceLast_ALL	1	-0.05747	0.00683	70.8166	<.0001	0.944
Had_NonClubOrder_Last1Months	1	-0.87379	0.17851	23.9612	<.0001	0.417
Avg_NonClub_ItemsPerOrder	1	0.03231	0.00742	18.9397	<.0001	1.033
STD_Cuml_All_DiscountOffers	1	0.36973	0.05025	54.1412	<.0001	1.447
STD_Cuml_All_ShippingOffers	1	-1.13816	0.09366	147.6678	<.0001	.
STD_Cuml_Club_Orders	1	4.77903	0.26190	332.9693	<.0001	118.989
STD_Cuml_ALL_Net	1	-0.00373	0.0005528	45.4218	<.0001	0.996
Avg_Club_ItemPrice_GT40	1	0.67529	0.12488	29.2424	<.0001	1.965
Cumu_ALL_DiscPct_GT20	1	-0.54452	0.09608	32.1182	<.0001	0.580
Cumu_Club_DiscPct_GT20	1	-0.53364	0.13185	16.3800	<.0001	0.586
STD_Cuml_Club_Net_GT40	1	0.76123	0.10441	53.1563	<.0001	2.141
Is_CoreClubMember	1	0.52418	0.14605	12.8808	0.0003	1.689
Last_ClubOrder_GT3months	1	1.09657	0.09934	121.8598	<.0001	2.994
STD_Cuml_All_ShippingOffers*STD_Cuml_All_ShippingOffers	1	0.27309	0.02468	122.4772	<.0001	.

Survival Analysis

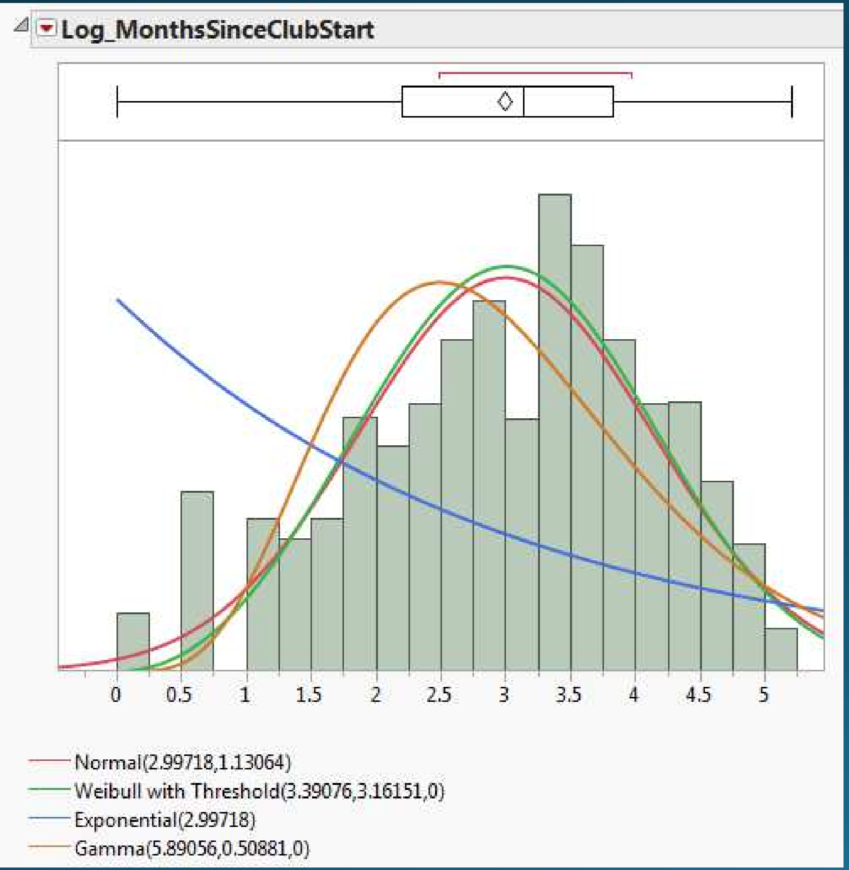
2. Semi-parametric model fit using PHREG – Winery B

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
Ever_Clicked	1	-0.60776	0.06395	90.3100	<.0001	0.545
Ever_Bounced	1	-0.47159	0.10075	21.9086	<.0001	0.624
MonthsSinceLast_ALL	1	-0.04438	0.00368	145.0643	<.0001	0.957
NonClubSale_Ever	1	0.51755	0.06276	68.0016	<.0001	1.678
Cumu_NonClub_DiscPct	1	0.02500	0.00657	14.4752	0.0001	1.025
Cumu_Club_DiscPct	1	-0.20420	0.01234	273.8574	<.0001	.
Moved_Last3Months	1	0.60414	0.16291	13.7532	0.0002	1.830
Recvd_Offer_Last1Months	1	-0.24320	0.06021	16.3141	<.0001	0.784
STD_Cuml_All_Opens	1	0.34481	0.05042	46.7702	<.0001	1.412
STD_Cuml_All_Clicks	1	1.66088	0.23638	49.3683	<.0001	.
STD_Cuml_All_Bounces	1	3.22136	0.59234	29.5763	<.0001	25.062
STD_Cuml_All_NoResponses	1	0.26742	0.04433	36.3852	<.0001	1.307
STD_Cuml_All_Reminders	1	-2.32337	0.29014	64.1229	<.0001	0.098
STD_Cuml_Club_Orders	1	0.81229	0.15089	28.9801	<.0001	2.253
STD_Cuml_Club_Disc	1	0.08520	0.01301	42.8991	<.0001	.
Cumu_Club_DiscPct_GT20	1	-2.28838	0.22969	99.2575	<.0001	0.101
Cumu_NonClub_DiscPct_GT0	1	-1.55506	0.12104	165.0602	<.0001	0.211
MilesFromWinery_GT100	1	0.33241	0.06172	29.0077	<.0001	1.394
Is_CoreClubMember	1	0.46935	0.07301	41.3306	<.0001	1.599
Last_ClubOrder_GT3months	1	0.54658	0.05985	83.3983	<.0001	1.727
STD_Cuml_All_Clicks*STD_Cuml_All_Clicks	1	-0.77186	0.15523	24.7259	<.0001	.
Cumu_Club_DiscPct*Cumu_Club_DiscPct	1	0.01220	0.0005537	485.6059	<.0001	.
STD_Cuml_Club_Disc*STD_Cuml_Club_Disc	1	-0.0008748	0.0001993	19.2658	<.0001	.

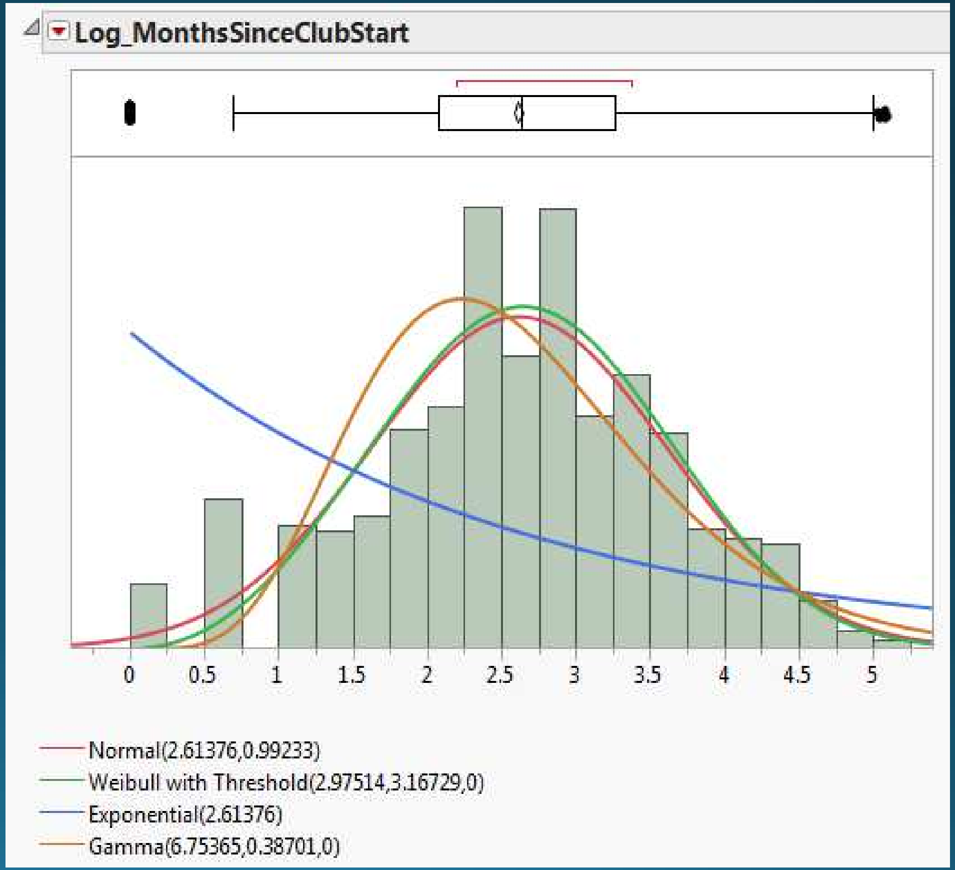
Survival Analysis

3. Used JMP to Evaluate Shape of Log(MonthsSinceClubStart)

Winery A



Winery B



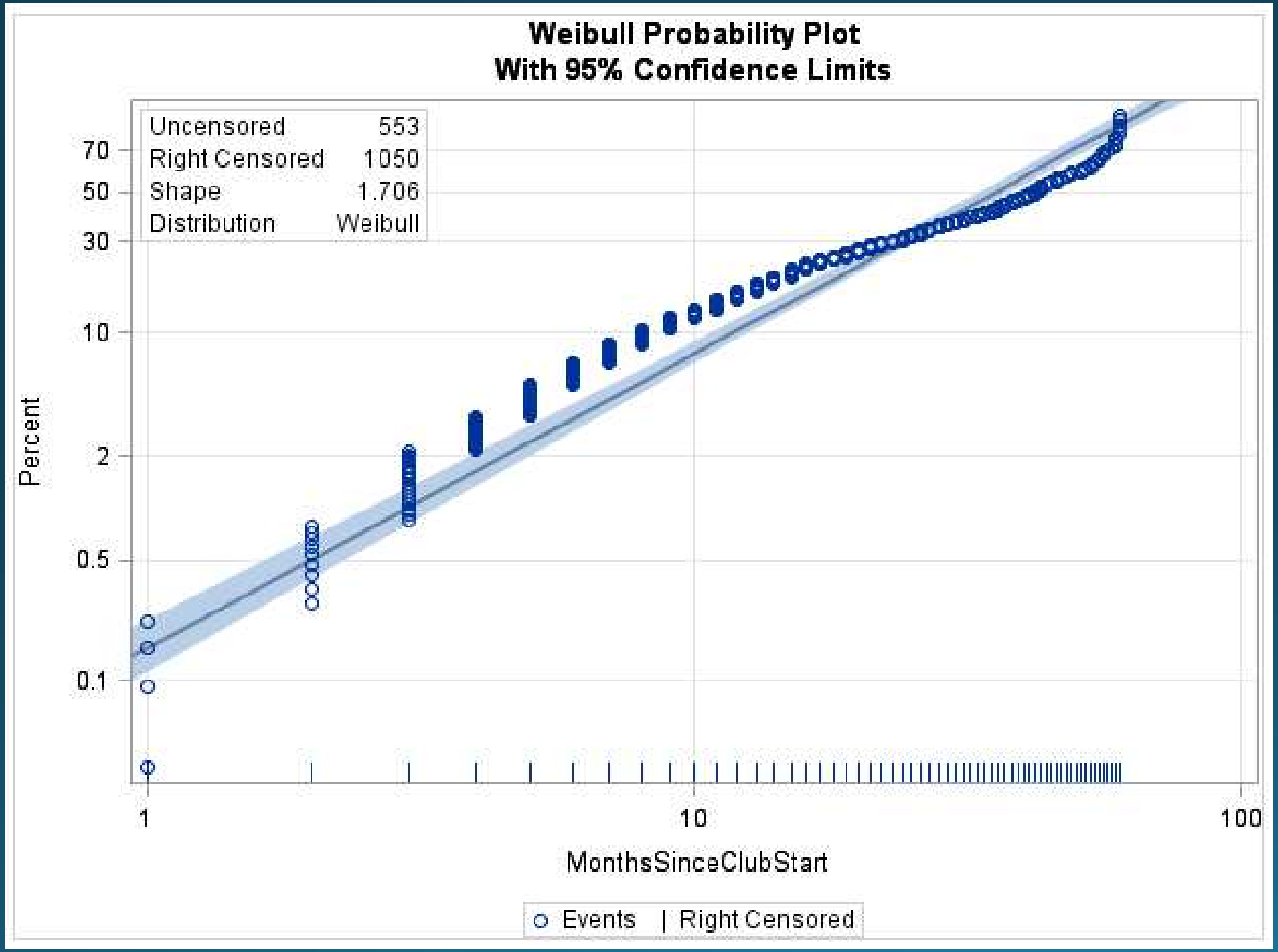
Survival Analysis

4. Training model fit using LIFEREG – Winery A

Analysis of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	4.7947	0.1431	4.5142	5.0752	1122.32	<.0001
MonthsSinceLast_Club	1	-0.0321	0.0088	-0.0494	-0.0149	13.30	0.0003
MonthsSinceLast_ALL	1	0.0323	0.0053	0.0219	0.0427	37.18	<.0001
Had_NonClubOrder_Las	1	0.4489	0.1307	0.1928	0.7051	11.80	0.0006
Avg_NonClub_ItemsPer	1	-0.0174	0.0058	-0.0288	-0.0061	9.06	0.0026
STD_Cuml_All_Discoun	1	-0.2071	0.0391	-0.2837	-0.1304	28.04	<.0001
STD_Cuml_All_Shippin	1	0.6933	0.0734	0.5494	0.8372	89.20	<.0001
STD_Cuml_Club_Orders	1	-2.6706	0.1928	-3.0485	-2.2927	191.88	<.0001
STD_Cuml_ALL_Net	1	0.0025	0.0004	0.0017	0.0033	39.71	<.0001
Avg_Club_ItemPrice_G	1	-0.4292	0.0937	-0.6128	-0.2455	20.98	<.0001
Cumu_ALL_DiscPct_GT2	1	0.3453	0.0708	0.2065	0.4841	23.79	<.0001
Cumu_Club_DiscPct_GT	1	0.2275	0.0969	0.0375	0.4175	5.51	0.0189
STD_Cuml_Club_Net_GT	1	-0.5176	0.0764	-0.6674	-0.3677	45.85	<.0001
Is_CoreClubMember	1	-0.2992	0.1064	-0.5077	-0.0907	7.91	0.0049
Last_ClubOrder_GT3mo	1	-0.6337	0.0812	-0.7930	-0.4745	60.84	<.0001
STD_Cuml_*STD_Cuml_A	1	-0.1593	0.0172	-0.1930	-0.1257	86.03	<.0001
Scale	1	0.5805	0.0183	0.5458	0.6174		
Weibull Shape	1	1.7227	0.0542	1.6197	1.8323		

Survival Analysis

4. Training model fit using LIFEREG – Winery A



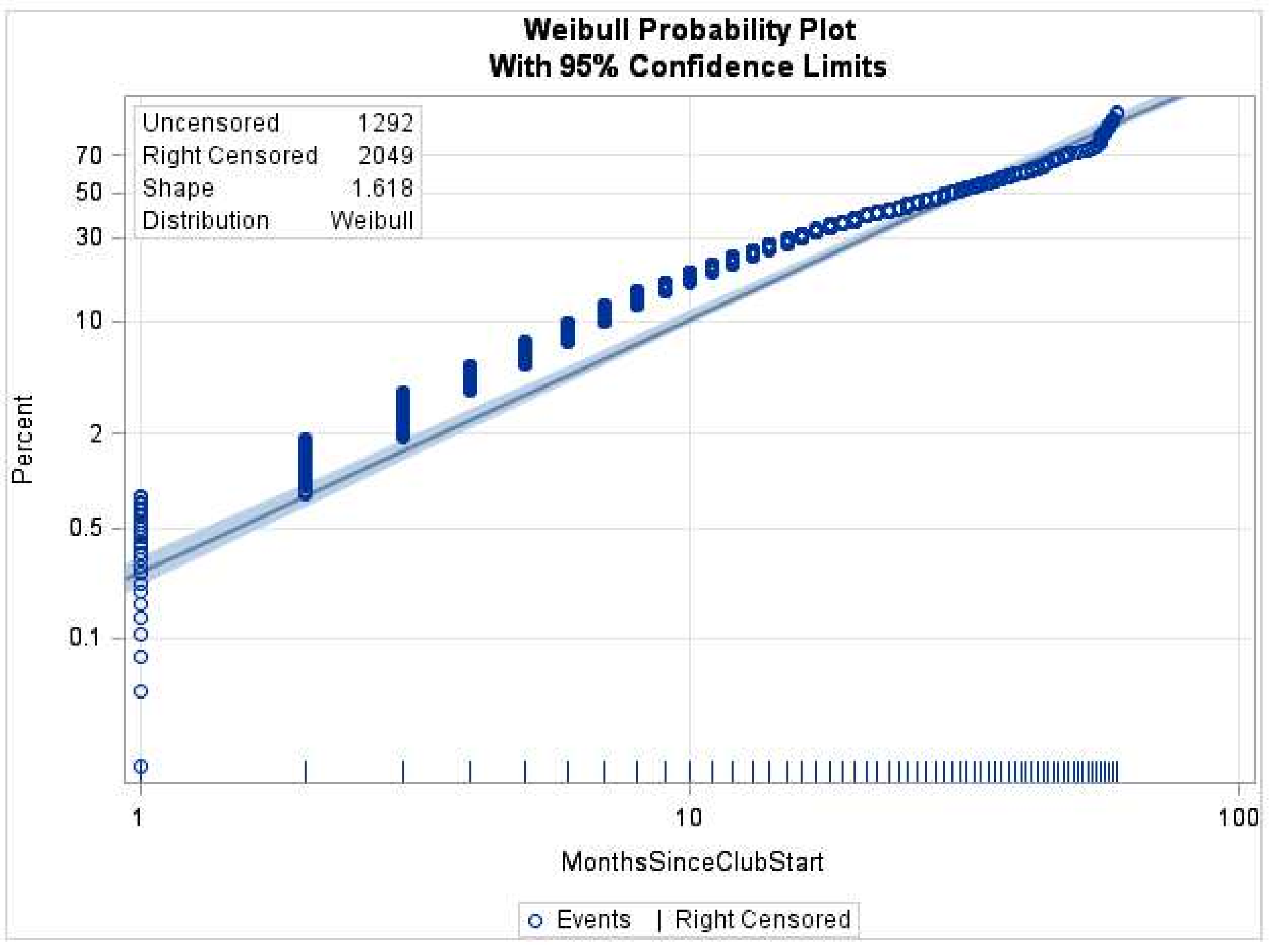
Survival Analysis

4. Training model fit using LIFEREG – Winery B

Analysis of Maximum Likelihood Parameter Estimates							
Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	3.7611	0.1251	3.5159	4.0062	904.21	<.0001
Ever_Clicked	1	0.3515	0.0522	0.2492	0.4538	45.36	<.0001
Ever_Bounced	1	0.3126	0.0837	0.1484	0.4767	13.93	0.0002
MonthsSinceLast_ALL	1	0.0301	0.0030	0.0243	0.0359	103.70	<.0001
NonClubSale_Ever	1	-0.3924	0.0497	-0.4899	-0.2949	62.24	<.0001
Cumu_NonClub_DiscPct	1	-0.0126	0.0051	-0.0226	-0.0025	5.98	0.0144
Cumu_Club_DiscPct	1	0.1435	0.0097	0.1244	0.1626	217.33	<.0001
Moved_Last3Months	1	-0.4188	0.1197	-0.6535	-0.1841	12.24	0.0005
Recvd_Offer_Last1Mon	1	0.1865	0.0484	0.0915	0.2814	14.82	0.0001
STD_Cuml_All_Opens	1	-0.2055	0.0411	-0.2861	-0.1250	25.03	<.0001
STD_Cuml_All_Clicks	1	-1.1978	0.1904	-1.5709	-0.8247	39.59	<.0001
STD_Cuml_All_Bounces	1	-2.3564	0.4572	-3.2525	-1.4602	26.56	<.0001
STD_Cuml_All_NoRespo	1	-0.1598	0.0368	-0.2320	-0.0877	18.85	<.0001
STD_Cuml_All_Reminde	1	1.2583	0.2332	0.8014	1.7153	29.13	<.0001
STD_Cuml_Club_Orders	1	-0.3707	0.1223	-0.6104	-0.1311	9.19	0.0024
STD_Cuml_Club_Disc	1	-0.0624	0.0122	-0.0862	-0.0386	26.33	<.0001
Cumu_Club_DiscPct_GT	1	1.3294	0.1972	0.9429	1.7159	45.45	<.0001
Cumu_NonClub_DiscPct	1	0.9554	0.0956	0.7681	1.1427	99.96	<.0001
MilesFromWinery_GT10	1	-0.1750	0.0488	-0.2707	-0.0793	12.85	0.0003
Is_CoreClubMember	1	-0.2903	0.0596	-0.4072	-0.1735	23.71	<.0001
Last_ClubOrder_GT3mo	1	-0.3101	0.0492	-0.4065	-0.2137	39.76	<.0001
STD_Cuml_*STD_Cuml_A	1	0.5973	0.1238	0.3547	0.8399	23.28	<.0001
Cumu_Club*Cumu_Club_	1	-0.0081	0.0004	-0.0090	-0.0073	350.66	<.0001
STD_Cuml_*STD_Cuml_C	1	0.0007	0.0002	0.0003	0.0012	9.34	0.0022
Scale	1	0.6181	0.0129	0.5933	0.6440		
Weibull Shape	1	1.6177	0.0339	1.5527	1.6855		

Survival Analysis

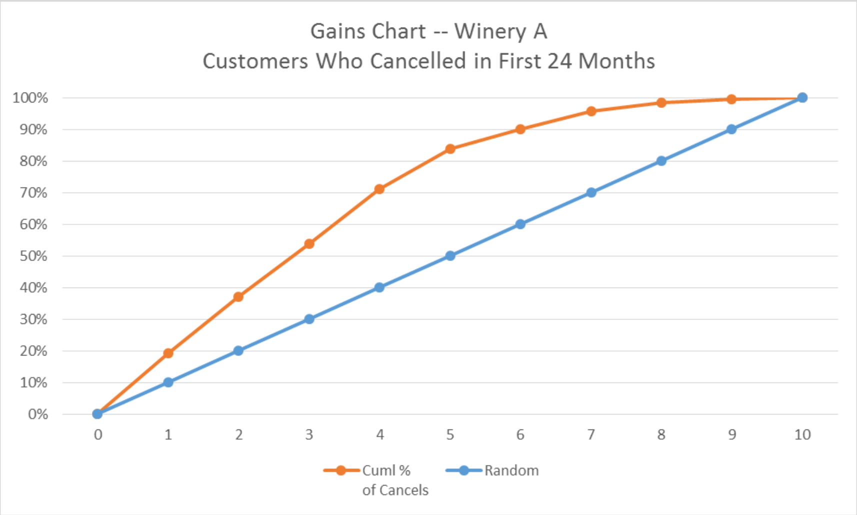
4. Training model fit using LIFEREG – Winery B



Survival Analysis

5. Validation model Gains Chart – Winery A

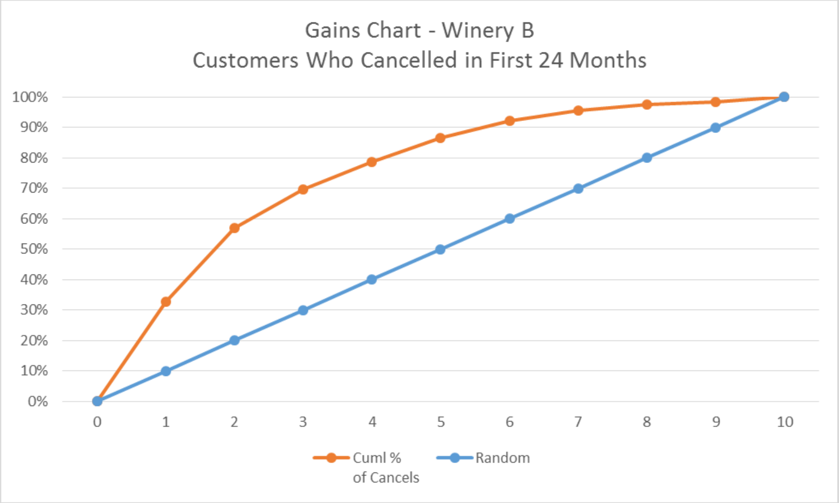
Decile	6 Months		12 Months		18 Months		24 Months		30 Months		36 Months	
	Cancels	Cuml % of Cancels	Cancels	Cuml % of Cancels	Cancels	Cuml % of Cancels	Cancels	Cuml % of Cancels	Cancels	Cuml % of Cancels	Cancels	Cuml % of Cancels
1	26	44.1%	33	25.6%	50	21.5%	49	19.3%	43	13.9%	38	11.7%
2	8	57.6%	34	51.9%	49	42.5%	45	37.0%	52	30.7%	54	28.4%
3	11	76.3%	24	70.5%	47	62.7%	43	53.9%	60	50.2%	63	47.8%
4	5	84.7%	14	81.4%	25	73.4%	44	71.3%	51	66.7%	58	65.7%
5	4	91.5%	11	89.9%	28	85.4%	32	83.9%	46	81.6%	53	82.1%
6	1	93.2%	3	92.2%	15	91.8%	16	90.2%	24	89.3%	17	87.3%
7	2	96.6%	6	96.9%	11	96.6%	14	95.7%	16	94.5%	16	92.3%
8	1	98.3%	2	98.4%	6	99.1%	7	98.4%	10	97.7%	18	97.8%
9	0	98.3%	0	98.4%	2	100.0%	3	99.6%	5	99.4%	4	99.1%
10	1	100.0%	2	100.0%	0	100.0%	1	100.0%	2	100.0%	3	100.0%



Survival Analysis

5. Validation model Gains Chart – Winery B

Decile	6 Months		12 Months		18 Months		24 Months		30 Months		36 Months	
	Cancels	Cuml % of Cancels	Cancels	Cuml % of Cancels	Cancels	Cuml % of Cancels	Cancels	Cuml % of Cancels	Cancels	Cuml % of Cancels	Cancels	Cuml % of Cancels
1	93	45.1%	173	39.1%	207	33.5%	219	32.6%	223	29.7%	219	28.7%
2	45	67.0%	101	61.9%	135	55.3%	163	56.9%	172	52.6%	156	49.2%
3	25	79.1%	54	74.0%	78	68.0%	85	69.6%	93	65.0%	93	61.4%
4	13	85.4%	40	83.1%	55	76.9%	61	78.7%	73	74.7%	80	71.9%
5	11	90.8%	28	89.4%	57	86.1%	52	86.4%	69	83.9%	72	81.4%
6	7	94.2%	22	94.4%	37	92.1%	38	92.1%	49	90.4%	47	87.5%
7	7	97.6%	10	96.6%	19	95.1%	24	95.7%	26	93.9%	43	93.2%
8	1	98.1%	11	99.1%	16	97.7%	13	97.6%	24	97.1%	25	96.5%
9	3	99.5%	1	99.3%	7	98.9%	6	98.5%	11	98.5%	23	99.5%
10	1	100.0%	3	100.0%	7	100.0%	10	100.0%	11	100.0%	4	100.0%



Survival Analysis

6. Validation Misclassification

Winery A

Cancellations Within 24 Months of Start Date
Posterior Probability: 0.36

Cutoff: 0.50	Predicted	
Actual	Active	Cancelled
Active	281	158
Cancelled	113	136

Misclass Rate: 39.39%
True Negative Rate: 64.01%
True Positive Rate: 54.62%

Cutoff: 0.40	Predicted	
Actual	Active	Cancelled
Active	257	182
Cancelled	79	170

Misclass Rate: 37.94%
True Negative Rate: 58.54%
True Positive Rate: 68.27%

Cutoff: 0.30	Predicted	
Actual	Active	Cancelled
Active	224	215
Cancelled	60	189

Misclass Rate: 39.97%
True Negative Rate: 51.03%
True Positive Rate: 75.90%

Winery B

Cancellations Within 24 Months of Start Date
Posterior Probability: 0.38

Cutoff: 0.50	Predicted	
Actual	Active	Cancelled
Active	946	142
Cancelled	205	466

Misclass Rate: 19.73%
True Negative Rate: 86.95%
True Positive Rate: 69.45%

Cutoff: 0.40	Predicted	
Actual	Active	Cancelled
Active	831	257
Cancelled	147	524

Misclass Rate: 22.97%
True Negative Rate: 76.38%
True Positive Rate: 78.09%

Cutoff: 0.30	Predicted	
Actual	Active	Cancelled
Active	667	421
Cancelled	82	589

Misclass Rate: 28.60%
True Negative Rate: 61.31%
True Positive Rate: 87.78%

Survival Analysis

7. Key Learnings

- It's difficult to get a great model fit. My theory is that this is due to the large number of censored observations however we also may not have the best predictors for this continuous outcome.
- The Winery A model fit is pretty bad. Perhaps this is due to a smaller dataset or significantly different underlying business processes than Winery B.
- In lieu of Survival Analysis, I think I would attempt to split the dataset into "early life" and "mature" customers and build separate logistic models.
- The underlying theory and assumptions of Survival Analysis are much more complex than Logistic or OLS. A great deal of study is likely required for this method to be optimized. Also, it would be pretty difficult to explain to a non-technical business manager.

Data Mining Analysis

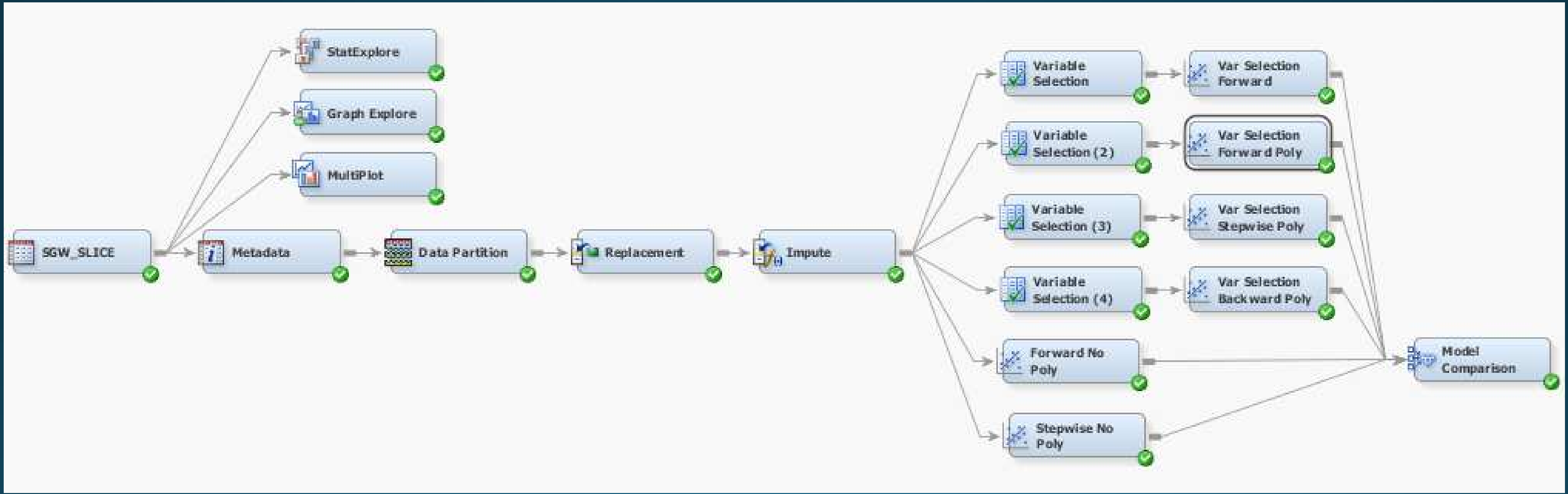
SAS Enterprise Miner (SEM) Analysis

1. Goals of Research

- Use a data mining approach to understand which logistic models perform best.
  - Provided SEM the main effects only.
  - Evaluated 6 stepwise options (SLENTER=0.10 / SLSTAY=0.05) and compared results:
    1. Variable Selection → Forward with NO interactions or quadratics
    2. Variable Selection → Forward WITH interactions and quadratics
    3. Variable Selection → Mixed WITH interactions and quadratics
    4. Variable Selection → Backward WITH interactions and quadratics
    5. NO Variable Selection → Forward WITH interactions and quadratics
    6. NO Variable Selection → Mixed WITH interactions and quadratics
- Evaluate the best logistic models to those created in JMP previously
  - Do the results look similar?
- Assess some of the tradeoffs between a data mining approach a more structured hypothesis-driven method

SAS Enterprise Miner Analysis

2. Flow Diagram



SAS Enterprise Miner (SEM) Analysis

3. Model Comparison – Winery A

Model Description	Selection Criterion: Valid: Misclassification Rate ▲	Train: Misclassification Rate	Valid: Average Squared Error	Train: Average Squared Error	Train: Akaike's Information Criterion
Var Selection Forward Poly	0.068838	0.055418	0.059501	0.043767	803.0277
Forward No Poly	0.105107	0.093518	0.08086	0.071105	1147.206
Var Selection Backward Poly	0.107328	0.017318	0.092929	0.015806	1077.202
Var Selection Forward	0.131014	0.107867	0.09507	0.082257	1221.2
Var Selection Stepwise Poly	0.310881	0.310242	0.214234	0.213992	2505.188
Stepwise No Poly	0.310881	0.310242	0.214234	0.213992	2505.188

SAS Enterprise Miner (SEM) Analysis

3. Best Performing Model – Winery A

- Validation Misclassification = 6.09%. 43 variables and 73 degrees of freedom

Effects	DF	Chi-Square	Pr > ChiSq
MonthsSinceFirst_EmailALL*STD_Cuml_All_ShippingOffers	1	97.7586	<.0001
MonthsSinceFirst_EmailALL*STD_Cuml_All_Sent	1	77.1963	<.0001
MonthsSinceFirst_ALL*STD_Cuml_All_Clicks	1	68.4516	<.0001
Had_ClubOrder_Last3Months	1	61.8101	<.0001
MonthsSinceFirst_EmailALL*MonthsSinceFirst_EmailALL	1	45.4793	<.0001
G_LengthOfResidence*G_MilesFromWineryGroup	20	37.3391	0.0107
Ever_Bounced*G_Division	4	32.1627	<.0001
Avg_Club_ItemPrice_GT40*STD_Cuml_All_DiscountOffers	1	28.9166	<.0001
MonthsSinceFirst_EmailALL*STD_Cuml_ALL_Net_GT100	1	26.1594	<.0001
MonthsSinceFirst_ALL*MonthsSinceFirst_ALL	1	23.7195	<.0001
STD_Cuml_All_ShippingOffers*STD_Cuml_All_ShippingOffers	1	22.0926	<.0001
STD_Cuml_All_DiscountOffers*STD_Cuml_All_DiscountOffers	1	21.7948	<.0001
MonthsSinceClubStart*STD_Cuml_Club_Orders	1	20.9372	<.0001
STD_Cuml_All_Reminders*STD_Cuml_Club_Orders	1	17.7775	<.0001
STD_Cuml_All_Sent*STD_Cuml_Club_Orders	1	16.6427	<.0001
MonthsSinceFirst_ALL*MonthsSinceLast_EmailALL	1	13.9176	0.0002
Avg_NonClub_ItemPrice*STD_Cuml_ALL_Net_GT100	1	13.7339	0.0002
Recvd_Offer_Last1Months*STD_Cuml_All_Reminders	1	13.0864	0.0003
Cumu_NonClub_DiscPct*Recvd_Offer_Last1Months	1	12.4817	0.0004
Avg_Club_ItemPrice	1	11.605	0.0007
G_ClubShipCarrier	4	10.9958	0.0266
Recvd_Offer_Last1Months*STD_Cuml_Club_Orders	1	10.8393	0.001

Effects (Continued)	DF	Chi-Square	Pr > ChiSq
Ever_Bounced*G_IncomeInd	4	10.4052	0.0341
MonthsSinceFirst_ALL*MonthsSinceFirst_EmailALL	1	10.3683	0.0013
MonthsSinceClubStart*MonthsSinceFirst_Club	1	10.2705	0.0014
Ever_Bounced*Had_ClubOrder_Last3Months	1	9.9542	0.0016
MonthsSinceLast_EmailALL*STD_Cuml_Club_Orders	1	9.4583	0.0021
MonthsSinceFirst_EmailALL*STD_Cuml_All_Reminders	1	8.2264	0.0041
G_LengthOfResidence*Had_ClubOrder_Last3Months	5	8.0715	0.1523
G_ClubSalesperson	6	7.2799	0.2957
Cumu_NonClub_DiscPct*MonthsSinceFirst_EmailALL	1	7.1501	0.0075
Avg_NonClub_ItemPrice*MonthsSinceLast_EmailALL	1	6.4456	0.0111
STD_Cuml_All_Sent*STD_Cuml_All_Sent	1	6.3259	0.0119
Avg_NonClub_ItemPrice*STD_Cuml_NonClub_Orders	1	5.3368	0.0209
STD_Cuml_All_DiscountOffers*STD_Cuml_All_ShippingOffers	1	4.774	0.0289
MonthsSinceClubStart*MonthsSinceFirst_NonClub	1	3.8394	0.0501
STD_Cuml_All_Opens	1	3.7558	0.0526
MonthsSinceFirst_EmailALL_GT24*MonthsSinceLast_EmailALL	1	2.9428	0.0863
Is_CoreClubMember*STD_Cuml_Club_Orders	1	2.0221	0.155
Recvd_Offer_Last1Months*STD_Cuml_All_Sent	1	0.0589	0.8083
MonthsSinceFirst_EmailALL*Recvd_Offer_Last1Months	1	0.0323	0.8574
IsClubOnHold*Recvd_Offer_Last1Months	1	0.0013	0.971
IsClubOnHold*Is_CoreClubMember	1	0.0004	0.984

SAS Enterprise Miner (SEM) Analysis

4. Model Comparison – Winery B

Model Description	Selection Criterion: Valid: Misclassification Rate ▲	Train: Misclassification Rate	Valid: Average Squared Error	Train: Average Squared Error	Train: Average Error Function	Train: Akaike's Information Criterion
Var Selection Forward Poly	0.060141	0.050416	0.049368	0.040256	0.153563	1218.725
Var Selection Backward Poly	0.065118	0.044875	0.056123	0.036226	0.139304	1167.778
Forward No Poly	0.075902	0.07036	0.058086	0.052591	0.192927	1456.931
Var Selection Forward	0.082124	0.076454	0.063964	0.058981	0.213414	1576.85
Var Selection Stepwise Poly	0.373289	0.37313	0.233944	0.233904	0.660601	4771.537
Stepwise No Poly	0.373289	0.37313	0.233944	0.233904	0.660601	4771.537

SAS Enterprise Miner (SEM) Analysis

4. Best Performing Model – Winery B

- Validation Misclassification = 6.0%. 50 variables and 54 degrees of freedom

Effect	DF	Chi-Square	Pr > ChiSq
STD_Cuml_All_Sent	1	69.8152	<.0001
STD_Cuml_All_DiscountOffers*STD_Cuml_All_ShippingOffers	1	38.5837	<.0001
Cumu_Club_DiscPct*STD_Cuml_All_DiscountOffers	1	34.895	<.0001
Avg_Club_SalesPerOrder_GT125*STD_Cuml_All_DiscountOffers	1	33.343	<.0001
STD_Cuml_All_DiscountOffers*STD_Cuml_All_DiscountOffers	1	28.7674	<.0001
Recvd_Offer_Last1Months	1	25.1157	<.0001
Avg_Club_SalesPerOrder*STD_Cuml_All_DiscountOffers	1	24.3505	<.0001
Avg_Club_ItemsPerOrder	1	23.5926	<.0001
Cumu_Club_DiscPct*STD_Cuml_Club_Disc	1	21.8532	<.0001
MonthsSinceFirst_ALL	1	21.2085	<.0001
Avg_Club_ItemPrice	1	20.1602	<.0001
MonthsSinceFirst_Club*STD_Cuml_All_Sent	1	17.515	<.0001
G_MilesFromWineryGroup	3	17.2818	0.0006
Cumu_Club_DiscPct	1	16.5861	<.0001
STD_Cuml_All_Reminders*STD_Cuml_All_Reminders	1	16.3244	<.0001
Avg_Club_ItemsPerOrder*STD_Cuml_All_DiscountOffers	1	15.7587	<.0001
Had_ClubOrder_Last3Months	1	15.7517	<.0001
Avg_Club_SalesPerOrder_GT125*MonthsSinceFirst_Club	1	15.4642	<.0001
Avg_Club_SalesPerOrder_GT125*STD_Cuml_All_Reminders	1	13.2366	0.0003
STD_Cuml_All_Sent*STD_Cuml_All_Sent	1	13.0164	0.0003
Avg_Club_ItemPrice*Cumu_Club_DiscPct	1	11.8225	0.0006
STD_Cuml_All_Reminders	1	11.7523	0.0006
Avg_Club_ItemPrice*MonthsSinceFirst_ALL	1	10.7568	0.001
MonthsSinceLast_EmailALL*STD_Cuml_All_ShippingOffers	1	10.5801	0.0011
Cumu_Club_DiscPct*Cumu_Club_DiscPct	1	10.5701	0.0011

Effect	DF	Chi-Square	Pr > ChiSq
STD_Cuml_Club_Items_GT1	1	10.225	0.0014
Avg_Club_ItemPrice*Avg_Club_SalesPerOrder	1	9.8088	0.0017
Cumu_Club_DiscPct*MonthsSinceFirst_Club	1	9.5098	0.002
Avg_Club_ItemPrice*MonthsSinceFirst_Club	1	9.3562	0.0022
MonthsSinceFirst_ALL*STD_Cuml_Club_Disc	1	8.5753	0.0034
STD_Cuml_All_Reminders*STD_Cuml_All_Sent	1	7.3414	0.0067
Avg_Club_ItemPrice*STD_Cuml_Club_Disc	1	7.2784	0.007
MonthsSinceFirst_Club*MonthsSinceFirst_Club	1	6.8596	0.0088
STD_Cuml_All_DiscountOffers*STD_Cuml_Club_Items_GT1	1	6.5662	0.0104
G_MilesFromWineryGroup*Recvd_Offer_Last1Months	3	6.5049	0.0895
STD_Cuml_All_Sent*STD_Cuml_Club_Items_GT1	1	6.2435	0.0125
MonthsSinceClubStart*MonthsSinceLast_EmailALL	1	6.1014	0.0135
STD_Cuml_Club_Disc	1	5.0624	0.0245
Cumu_NonClub_DiscPct_GT0*STD_Cuml_Club_Items_GT1	1	4.7454	0.0294
Avg_Club_SalesPerOrder_GT125*MonthsSinceClubStart	1	4.1683	0.0412
Cumu_Club_DiscPct*STD_Cuml_Club_Items_GT1	1	3.9705	0.0463
STD_Cuml_All_ShippingOffers*STD_Cuml_All_ShippingOffers	1	3.5818	0.0584
MonthsSinceFirst_ALL*MonthsSinceFirst_ALL	1	3.5107	0.061
Avg_Club_ItemsPerOrder*STD_Cuml_Club_Items_GT1	1	3.348	0.0673
STD_Cuml_All_DiscountOffers*STD_Cuml_All_Reminders	1	1.6363	0.2008
STD_Cuml_All_Reminders*STD_Cuml_Club_Items_GT1	1	0.6428	0.4227
Cumu_NonClub_DiscPct_GT0	1	0.3626	0.5471
Cumu_NonClub_DiscPct_GT0*MonthsSinceLast_EmailALL	1	0.0961	0.7566
Avg_Club_SalesPerOrder*STD_Cuml_All_Reminders	1	0.0642	0.8
STD_Cuml_All_Reminders*STD_Cuml_All_ShippingOffers	1	0.0331	0.8556

SAS Enterprise Miner (SEM) Analysis

7. Key Learnings

- SAS Enterprise Miner provides a great graphical user interface to do sophisticated data mining task and can generate results equal or better than traditional methods.
- A drawback is that there is very little emphasis on reports and plots that can confirm if the model is correctly specified. One could use the SAS node within SEM or use SAS outside of SEM to write code to assess model validity.
- In this example, SEM was very efficient at testing many different interactions and quadratics and was more than willing to use these liberally. The result was a very high percentage of terms in the final model being quadratics of some sort. The models were considerably bigger than the model identified through traditional methods.
- For very large datasets where predictive power is of higher importance than understanding underlying associations, SEM really excels. However, the models may be overly dimensional and need to be retrained often to maintain results.